Information Acquisition Under Resource Limitations in a Noisy Environment

Matvey Soloviev
Computer Science Department
Cornell University
msoloviev@cs.cornell.edu

Joseph Y. Halpern
Computer Science Department
Cornell University
halpern@cs.cornell.edu

Abstract

We introduce a theoretical model of information acquisition under resource limitations in a noisy environment. An agent must guess the truth value of a given Boolean formula φ after performing a bounded number of noisy tests of the truth values of variables in the formula. We observe that, in general, the problem of finding an optimal testing strategy for φ is hard, but we suggest a useful heuristic. The techniques we use also give insight into two apparently unrelated, but well-studied problems: (1) *rational inattention* (the optimal strategy may involve hardly ever testing variables that are clearly relevant to φ) and (2) what makes a formula hard to learn/remember.

1 Introduction

Decision-making is typically subject to resource constraints. However, an agent may be able to choose how to allocate his resources. We consider a simple decision-theoretic framework in which to examine this resource-allocation problem. To motivate the framework, consider an animal that must decide whether some food is safe to eat. We assume that "safe" is characterised by a known Boolean formula φ, which depends on pertinent variables such as presence of unusual smells or signs of other animals consuming the same food. The animal can perform a limited number of tests of the variables in φ, but these tests are noisy; if a test says that a variable v is true, that does not mean that v is true, but only that it is true with some probability. After the agent has exhausted his test budget, he must either guess the truth value of φ or choose not to guess. Depending on his choice, he gets a payoff. In this example, guessing that φ is true amounts to guessing that the food is safe to eat. There will be a small positive payoff for guessing "true" if the food is indeed safe, but a large negative payoff for guessing "true" if the food is not safe to eat. In this example we can assume a payoff of 0 if the agent guesses "false" or does not guess, since both choices amount to not eating the food.

We are interested in optimal strategies for this decision; that is, what tests should the agent perform and in what order. Unfortunately (and perhaps not surprisingly), as we show, finding an optimal strategy (i.e., one that obtains the highest expected payoff) is infeasibly hard. We provide a heuristic that guarantees a positive expected payoff whenever the optimal strategy gets a positive expected payoff. Our analysis of this strategy also gives us the tools to examine two other problems of interest.

The first is *rational inattention*, the notion that in the face of limited resources it is sometimes rational to ignore certain sources of information completely. There has been a great deal of interest recently in this topic in economics [Sims, 2003; Wiederholt, 2010]. Here we show that optimal testing strategies in our framework exhibit what can reasonably be called rational inattention (which we typically denote RI from now on). Specifically, our experiments show that for a substantial fraction of formulae, an optimal strategy will hardly ever test variables that are clearly relevant to the outcome. (Roughy speaking, "hardly ever" means that as the total number of tests goes to infinity, the fraction of tests devoted to these relevant variables goes to 0.) For example, consider the formula $v_1 \vee v_2$. Suppose that the tests for v_1 and v_2 are equally noisy, so there is no reason to prefer one to the other for the first test. But for appropriate choices of payoffs, we show that if we start by testing v_2, then all subsequent tests should also test v_2 as long as v_2 is observed to be true (and similarly for v_1). Thus, with positive probability, the optimal strategy either ignores v_1 or ignores v_2. Our formal analysis allows us to conclude that this is a widespread phenomenon.

The second problem we consider is what makes a concept (which we can think of as being characterised by a formula) hard. To address this, we use our framework to define a notion of hardness. We show that,

according to this definition, XORs (i.e., formulae of the form $v_1 \oplus \cdots \oplus v_n$, which are true exactly if an odd number of the v_i's are true) and their negations are the hardest formulae. We compare this notion to other notions of hardness of concepts considered in the cognitive psychology literature (e.g., [Feldman, 2006; Love, Medin, and Gureckis, 2004; Shepard, Hovland, and Jenkins, 1961]).

2 Information-acquisition games

We model the *information-acquisition game* as a single-player game against nature. The game is characterised by five parameters:

- a Boolean formula φ that mentions variables $v_1, \ldots, v_n$ for some $n > 0$;

- a probability distribution D on truth assignments to $\{v_1, \ldots, v_n\}$;
- a bound k on the number of tests;
- an *accuracy vector* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)$, with $0 \leq \alpha_i \leq 1/2$ (explained below);
- payoffs (g, b), where $g > 0 > b$ (also explained below).

We denote this game as $G(\varphi, D, k, \vec{\alpha}, g, b)$.

In the game $G(\varphi, D, k, \vec{\alpha}, g, b)$, nature first chooses a truth assignment to the variables $v_1, \ldots, v_n$ according to distribution D. While the parameters of the game are known to the agent, the assignment chosen by nature is not. For the next k rounds, the agent then chooses one of the n variables to test (perhaps as a function of history), and nature responds with either T or F. The agent then must either guess the truth value of φ or choose not to guess. We view a truth assignment as a function from variables to truth values ($\{T, F\}$); we can also view the formula φ itself as a function from truth assignments to truth values. If the agent chooses to test v_i, then nature returns $A(v_i)$ (the right answer) with probability $1/2 + \alpha_i$ (and thus returns $\neg A(v_i)$ with probability $1/2 - \alpha_i$).[1] Finally, if the agent choses not to guess at the end of the game, his payoff is 0. If he chooses to guess, then his payoff is g (good) if he guesses $\varphi(A)$ (i.e., if he guesses the truth value of φ correctly, given that A is the actual truth assignment) and b (bad) if he guesses $\neg\varphi(A)$. A strategy for an agent in this game is just a function that determines which test the agent performs after observing each test-outcome sequence of length $< k$, together with a final action for each test-outcomes sequence of length k.

Example 2.1. Consider the information-acquisition game over the formula $v_1 \vee v_2$, with $k = 2$ tests, a uniform distribution on truth assignments, accuracy vector $(1/4, 1/4)$, correct-guess reward $g = 1$ and wrong-guess penalty $b = -16$. As we show (see Appendix A) this game has two optimal strategies:

1. test v_1 twice, guess T if both tests came out T, and make no guess otherwise;
2. test v_2 twice, guess T if both tests came out T, and make no guess otherwise.

□

Thus, in this game, an optimal strategy either ignores v_1 or ignores v_2. As we show in the full paper, the strategy "test v_1 and then v_2, guess T if both tests came out T" is strictly worse than these two; in fact, its expected payoff is negative!

If we increase k, the situation becomes more nuanced. For instance, if $k = 4$, an optimal strategy tests v_1 once, and if the test comes out F, tests v_2 three times and guesses T if all three tests came out T. However, it always remains optimal to keep testing one variable as long as the tests keep coming out true. That is, all optimal strategies exhibit RI in the sense that there are test outcomes that result in either v_1 never being tested or v_2 never being tested, despite their obvious relevance to $v_1 \vee v_2$.

We will frequently talk about the probability of various events over the course of a run of the game, and many of the probabilities we care about depend only on a few parameters of the game. Formally, we embed the traces of all information-acquisition games on formulae in n variables in a common sample space Ω_n, whose elements are tuples of the form

$$(A, v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k, a),$$

where A is the assignment of truth values to the n variables chosen by nature, $v_{i_j} \approx b_j$ indicates that the jth test was performed on variable v_{i_j} and that nature responded with the test outcome b_j, and a is the final agent action of either making no guess or guessing some truth value for the formula. A game $G(\varphi, D, k, \vec{\alpha}, g, b)$ and agent strategy σ then induce a probability $\Pr_{G,\sigma}$ on this sample space. The only features of the game G that affect the probability are the prior distribution D and the accuracy vector α, so we write $\Pr_{D,\alpha,\sigma}(\varphi)$ rather than $\Pr_{G,\sigma}(\varphi)$. If some component of the subscript does not affect the probability, then we typically omit it. In particular, we show in Appendix B that the strategy σ does not affect $\Pr_{G,\sigma}(\varphi|S)$, so we write $\Pr_{D,\vec{\alpha}}(\varphi|S)$. Finally, the utility (payoff) received by the agent at the end of the game is a real-valued random variable that depends on parameters b and g. We can define the expected utility $\mathbb{E}_{G,\sigma}(\text{payoff})$ as the expectation of this random variable.

3 Determining optimal strategies

It is straightforward to see that the game tree for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$ has $3(2^n)(2n)^k$ leaves: there is a branching factor of 2^n at the root (since there are 2^n truth assignments) followed by k branching factors of n (for the n variables that the agent can choose to test) and 2 (for the two possible outcomes of a test). At the end there are three choices (don't guess, guess T, and guess F). A straightforward backward induction can then be used to compute the optimal strategy. Unfortunately, the complexity of this approach is polynomial in the number of leaves; it quickly grows infeasible as k grows.

In general, it is unlikely that the dependency on 2^n can be removed. In the special case that $b = -\infty$ and $\alpha_i = \frac{1}{2}$ for all i (so tests are perfectly accurate, but the truth value of the formula must be established for sure), determining whether a strategy of length k gets a positive expected payoff reduces to the problem of finding a conjunction of length k that implies a given Boolean formula. Umans [1999] showed that this problem is Σ_2^p-complete, that is, lies in a complexity class that is at least as hard as both NP and co-NP.

A simple heuristic (that is independent of φ) would simply be to test each variable in φ k/n times, and then choose the action that maximises the expected payoff given the observed test outcomes. We can calculate in time polynomial in k and n the expected payoff of a guess, conditional on a sequence of test outcomes. Since determining the best guess involves checking the likelihood of each of the 2^n truth assignments conditional on the outcomes, this approach takes time polynomial in k and 2^n. We are most interested in formulae where n

[1] Note that this means that the probability of a false positive and that of a false negative are the same. While we could easily extend the framework so as to allow the accuracy in a test on a variable v to depend on whether $A(v)$ is T or F, doing so would complicate notation and distract from the main points that we want to make.

is small, so this time complexity would be acceptable. However, this approach can be arbitrarily worse than the optimum. As we observed in Example 2.1, the expected payoff of this strategy is negative, while there is a strategy that has positive expected payoff.

An arguably somewhat better heuristic, which we call the *random-test heuristic*, is to choose, at every step, the next variable to test uniformly at random, and again, after k observations, choosing the action that maximises the expected payoff. This heuristic clearly has the same time complexity as the preceding one, while working better in information-acquisition games that require an unbalanced approach to testing.

Proposition 3.1. *If there exists a strategy that has positive expected payoff in the information-acquisition game G, then the random-test heuristic has positive expected payoff.*

To prove Proposition 3.1, we need a preliminary lemma. Intuitively, an optimal strategy should try to generate test-outcome sequences S that maximise $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$, since the larger $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ is, the more certain the agent is regarding whether φ is true or false. The following lemma characterises how large $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ has to be to get a positive expected payoff.

Definition 3.2. Let $q(b,g) = \frac{b+g}{2(b-g)}$ be the *threshold* associated with payoffs b, g. □

Lemma 3.3. *The expected payoff of $G(\varphi, D, k, \vec{\alpha}, g, b)$ when making a guess after observing a sequence S of test outcomes is positive if and only if*

$$|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2| > q(b,g). \tag{1}$$

Proof. The expected payoff when guessing that the formula is true is

$$g \cdot \Pr_{D,\vec{\alpha}}(\varphi \mid S) + b \cdot (1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)).$$

This is greater than zero iff

$$(g-b)\Pr_{D,\vec{\alpha}}(\varphi \mid S) + b > 0,$$

that is, iff

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2 > \frac{b}{b-g} - \frac{1}{2} = q(b,g).$$

When guessing that the formula is false, we simply exchange $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ and $1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)$ in the derivation. So the payoff is then positive iff

$$(1 - \Pr_{D,\vec{\alpha}}(\varphi \mid S)) - \frac{1}{2} = -(\Pr_{D,\vec{\alpha}}(\varphi \mid S) - \frac{1}{2}) > q(b,g).$$

Since $|x| = \max\{x, -x\}$, at least one of these two inequalities must hold if (1) does, so the corresponding guess will have positive expected payoff. Conversely, since $|x| \geq x$, either inequality holding implies (1). □

Proof of Proposition 3.1. Suppose that σ is a strategy for G with positive expected payoff. The outcome sequences of length k partition the space of paths in the game tree, so we have

$$\sum_{\{S:|S|=k\}} \Pr_{D,\vec{\alpha},\sigma}(S)\mathbb{E}_{G,\sigma}(\text{payoff} \mid S).$$

Since the payoff is positive, at least one of the summands on the right must be, say the one due to the sequence S^*. By Lemma 3.3, $|\Pr_{D,\vec{\alpha}}(\varphi \text{ is true} \mid S^*) - 1/2| > q(b,g)$.

Let τ denote the random-test heuristic. Since τ chooses the optimal action after making k observations, it will not get a negative expected payoff for any sequence S of k test outcomes (since it can always obtain a payoff of 0 by choosing not to guess). On the other hand, with positive probability, the variables that make up the sequence S^* will be chosen and the outcomes in S^* will be observed for these tests; that is $\Pr_{D,\vec{\alpha},\tau}(S^*) > 0$. It follows from Lemma 3.3 that $\mathbb{E}_{G,\tau}(\text{payoff} \mid S^*) > 0$. Thus, $\mathbb{E}_{G,\tau}(\text{payoff}) > 0$, as desired. □

4 Rational inattention

We might think that an optimal strategy for learning about φ would test all variables that are relevant to φ (given a sufficiently large test budget). As shown in Example 2.1, this may not be true. For example, an optimal k-step strategy for $v_1 \vee v_2$ can end up never testing v_1, no matter what the value of k, if it starts by testing v_2 and keeps discovering that v_2 is true. It turns out that RI is quite widespread.

It certainly is not surprising that if a variable v does not occur in φ, then an optimal strategy would not test v. More generally, it would not be surprising that a variable that is not particularly relevant to φ is not tested too often, perhaps because it makes a difference only in rare edge cases. In the foraging animal example from the introduction, the possibility of a human experimenter having prepared a safe food to look like a known poisonous plant would impact whether it is safe to eat, but is unlikely to play a significant role in day-to-day foraging strategies. What might seem more surprising is if a variable v is (largely) ignored while another variable v' that is no more relevant than v is tested. This is what happens in Example 2.1; although we have not yet defined a notion of relevance, symmetry considerations dictate that v_1 and v_2 are equally relevant to $v_1 \vee v_2$, yet an optimal strategy might ignore one of them.

The phenomenon of rational inattention observed in Example 2.1 is surprisingly widespread. To make this claim precise, we need to define "relevance". There are a number of reasonable ways of defining it; we focus on one below, although our results hold for other reasonable definitions too. The definition of the relevance of v to φ that we use counts the number of truth assignments for which changing the truth value of v changes the truth value of φ.

Definition 4.1. Define the relevance ordering $\leq_\varphi$ on the variables in φ by taking

$$\begin{aligned} & v \leq_\varphi v' \text{ iff} \\ & |\{A : \varphi(A[v \mapsto \mathrm{T}]) \neq \varphi(A[v \mapsto \mathrm{F}])\}| \\ \leq\ & |\{A : \varphi(A[v' \mapsto \mathrm{T}]) \neq \varphi(A[v' \mapsto \mathrm{F}])\}|, \end{aligned}$$

where $A[v \mapsto b]$ is the assignment that agrees with A except that it assigns truth value b to v. □

Thus, rather than saying that v is or is not relevant to φ, we can say that v is (or is not) at least as relevant to φ as v'. Considering the impact of a change in a single variable to

the truth value of the whole formula in this fashion has been done both in the cognitive science and the computer science literature: for example, Vigo [2011] uses the *discrete (partial) derivative* to capture this effect, and Lang et al. [2003] define the related notion of *Var-independence*.

We could also consider taking the probability of the set of truth assignments where a variable's value makes a difference, rather than just counting how many such truth assignments there are. This would give a more detailed quantitative view of relevance, and is essentially how relevance is considered in Bayesian networks. Irrelevance is typically identified with independence. Thus, v is relevant to φ if a change to v changes the probability of φ. (See Druzdzel and Suermondt [1994] for a review of work on relevance in the context of Bayesian networks.) We did not consider a probabilistic notion of relevance because then the relevance order would depend on the game (specifically, the distribution D, which is one of the parameters of the game). Our definition makes the relevance order depend only on φ. That said, we believe that essentially the same results as those we prove would obtain for a probabilistic notion of relevance ordering.

Roughly speaking, φ exhibits RI if, for all optimal strategies σ for the game $G(\varphi, D, k, \vec{\alpha}, b, g)$, with probability 1, σ tests a variable v' frequently while hardly ever testing a variable v that is at least as relevant to φ as v'. We still have to make precise "hardly ever", and explain how the claim depends on the choice of D, $\vec{\alpha}$, k, b, and g. For the latter point, note that in Example 2.1, we had to choose b and g appropriately to get RI. This turns out to be true in general; given D, k, and $\vec{\alpha}$, the claim holds only for an appropriate choice of b and g that depends on these. In particular, for any fixed choice of b and g that depends only on k and $\vec{\alpha}$, there exist choices of priors D for which the set of optimal strategies is fundamentally uninteresting: we can simply set D to assign a probability to some truth assignment A that is so high that no k test outcomes could make it rational to make any other guess than that the truth value of the formula is $\varphi(A)$.

Another way that the set of optimal strategies can be rendered uninteresting is when, from the outset, there is no hope of obtaining sufficient certainty of the formula's truth value with the k tests available. Similarly to when the truth value is a foregone conclusion, in this situation, an optimal strategy can perform arbitrary tests, as long as it makes no guess at the end. More generally, even when in general the choice of variables to test does matter, a strategy can reach a situation where there is sufficient uncertainty that no future test outcome could affect the final choice. Thus, a meaningful definition of RI that is based on the variables tested by optimal strategies must consider only tests performed in those cases in which a guess actually should be made (because the expected payoff of the optimal strategy is positive).[2] We now make these ideas precise.

[2]One way to avoid these additional requirements is to modify the game so that performing a test is associated has a small but positive cost, so that an optimal strategy avoids frivolous testing when the conclusion is foregone. The definitions we use have essentially the same effect, and are easier to work with.

Definition 4.2. A function $f : \mathbb{N} \to \mathbb{N}$ is *negligible* if $\lim_{k\to\infty} f(k)/k = 0$. □

The idea is that φ exhibits RI if, as the number k of tests allowed increases, the fraction of times that some variable v is tested is negligible relative to the number of times that another variable v' is tested, although v is at least as relevant to φ as v'. We actually require slightly more: we want v' to be tested a linear number of times (i.e., at least ck times, for some constant $c > 0$).

Since we do not want our results to depend on correlations between variables, we restrict attention to probability distributions D on truth assignments that are product distributions.

Definition 4.3. A probability distribution D on truth assignments to n variables is a *product distribution* if there exist probability distributions D_i on truth assignments to v_i for $i = 1, \ldots, n$ such that $D = D_1 \times \cdots \times D_n$. □

As discussed earlier, to get an interesting notion of RI, we need to allow the choice of payoffs b and g to depend on the prior distribution D; for fixed b, g, and testing bound k, if the distribution D places sufficiently high probability on a single assignment, no k outcomes can change the agent's mind. For similar reasons, we do not want to allow D to give a single assignment probability 1. More generally, assigning prior probability 1 to any one variable being true or false means that no tests will change the agent's mind about that variable, and so testing it is pointless (and the game is therefore equivalent to one played on the formula in $n - 1$ variables where this variable has been replaced by the appropriate truth value). We say that a probability distribution that gives all truth assignments positive probability is *open-minded*.

With all these considerations in hand, we can finally define RI formally.

Definition 4.4. The formula φ *exhibits rational inattention* if, for all open-minded product distributions D and uniform accuracy vectors $\vec{\alpha}$ (those with $(\alpha_1 = \ldots = \alpha_n)$), there exists a negligible function f and a constant $c > 0$ such that for all k, there are payoffs b and g such that all optimal strategies in the information-acquisition game $G(\varphi, D, k, \vec{\alpha}, b, g)$ have positive expected payoff and, in all runs of the game, depending on outcomes of tests, either make no guess or

- test a variable v' at least ck times, but
- test a variable v such that $v' \leq_\varphi v$ at most $f(k)$ times.

□

We can check in a straightforward way whether some natural classes of formulae exhibit RI in the sense of this definition.

Example 4.5. (Rational inattention)

1. Conjunctions $\varphi = \bigwedge_{i=1}^{N} \ell_i$ and disjunctions $\varphi = \bigvee_{i=1}^{N} \ell_i$ of $N \geq 2$ literals (variables $\ell_i = v_i$ or their negations $\neg v_i$) exhibit RI. In each case, we can pick b and g such that all optimal strategies pick one variable and focus on it, either to establish that the formula is true (for disjunctions) or that it is false (for conjunctions). By symmetry, all variables v_i and v_j are equally relevant, so $v_i \leq_\varphi v_j$.

2. The formulae v_i and $\neg v_i$ do not exhibit RI. There is no variable $v \neq v_i$ such that $v_i \leq_{(\neg)v_i} v$, and for all choices of b and g, the strategy of testing only v_i and ignoring all other variables (making an appropriate guess in the end) is clearly optimal for $(\neg)v_i$.
3. More generally, we can say that all XORs in ≥ 0 variables do not exhibit RI. For the constant formulae T and F, any testing strategy that "guesses" correctly is optimal; for any XOR in more than one variable, an optimal strategy must test all of them as any remaining uncertainty about the truth value of some variable leads to at least equally great uncertainty about the truth value of the whole formula. Similarly, negations of XORs do not exhibit RI. Together with the preceding two points, this means that the only formulae in 2 variables exhibiting rational inattention are the four conjunctions $\ell_1 \wedge \ell_2$ and the four disjunctions $\ell_1 \vee \ell_2$ in which each variable occurs exactly once and may or may not be negated.
4. For $n > 2$, formulae φ of the form $v_1 \vee (\neg v_1 \wedge v_2 \wedge \ldots \wedge v_n))$ do not exhibit RI. Optimal strategies that can attain a positive payoff at all will start by testing v_1; if the tests come out true, it will be optimal to continue testing v_1, ignoring $v_2 \ldots v_n$. However, for formulae φ of this form, v_1 is strictly more relevant than the other variables: there are only 2 assignments where changing v_i flips the truth value of the formula for $i > 1$ (the two where $v_1 \mapsto F$ and $v_j \mapsto T$ for $j \notin \{1, i\}$) but $2^n - 2$ assignments where changing v_1 does (all but the two where $v_j \mapsto T$ for $j \neq 1$). Hence, in the event that all these tests actually succeed, the only variables that are ignored are not at least as relevant as the only one that isn't, so φ does not exhibit RI.

□

Unfortunately, as far as we know, determining the optimal strategies is hard in general. To be able to reason about whether φ exhibits RI in a tractable way, we find it useful to consider optimal test-outcome sequences.

Definition 4.6. A sequence S of test outcomes is *optimal* for a formula φ, prior D, and accuracy vector $\vec{\alpha}$ if it minimises the conditional uncertainty about the truth value of φ among all test-outcome sequences of the same length. That is,

$$\left|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - \frac{1}{2}\right| \geq \left|\Pr_{D,\vec{\alpha}}(\varphi \mid S') - \frac{1}{2}\right|$$

for all S' with $|S'| = |S|$. □

Using this definition, we can derive a sufficient (but not necessary!) condition for formulae to exhibit RI.

Proposition 4.7. *Suppose that, for a given formula φ, for all open-minded product distributions D and uniform accuracy vectors $\vec{\alpha}$, there exists a negligible function f and a constant $c > 0$ such that for all testing bounds k, the test-outcome sequences S optimal for φ, D, and $\vec{\alpha}$ of length k have the following two properties:*

- *S has at least ck tests of some variable v', but*
- *S has at most $f(k)$ tests of some variable $v \geq_\varphi v'$.*

Then φ exhibits RI.

Proof. Let $P(\varphi, D, \vec{\alpha}, f, c, k)$ denote the statement that for all test-outcomes sequences S that are optimal for φ, D, and $\vec{\alpha}$, there exist variables $v \geq_\varphi v'$ such that S contains $\geq ck$ tests of v' and $\leq f(k)$ tests of v. We now prove that for all φ, D, $\vec{\alpha}$, f, c, and k, $P(\varphi, D, \vec{\alpha}, f, c, k)$ implies the existence of b and g such that φ exhibits RI in the game $G(\varphi, D, k, m, b, g)$. It is easy to see that this suffices to prove the proposition.

Fix φ, D, $\vec{\alpha}$, f, c, and k, and suppose that $P(\varphi, D, \vec{\alpha}, f, c, k)$ holds. Let

$$q^* = \max_{\{S:|S|=k\}} \left|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}\right|.$$

Since there are only finitely many outcome sequences of length k, there must be some $\epsilon > 0$ sufficiently small such that for all S with $|S| = k$, $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| > q^* - \epsilon$ if and only if $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| = q^*$. Choose the payoffs b and g such that the threshold $q(b, g) = q^* - \epsilon$. We show that φ exhibits RI in the game $G(\varphi, D, k, m, b, g)$.

Let $\mathcal{S}_k = \{S : |S| = k \text{ and } |\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| = q^*\}$ be the set of test-outcome sequences of length k optimal for φ, D, and $\vec{\alpha}$. If σ is an optimal strategy for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$, the only sequences of test outcomes after which σ makes a guess are the ones in $\mathcal{S}_k$. For if a guess is made after seeing some test-outcome sequence $S^* \notin \mathcal{S}_k$, by Lemma 3.3 and the choice of b and g, that expected payoff of doing so must be negative, so the strategy σ' that is identical to σ except that it makes no guess if S^* is observed is strictly better than σ, contradicting the optimality of σ. So whenever a guess is made, it must be after a sequence $S \in \mathcal{S}_k$ was observed. Since sequences in $\mathcal{S}_k$ are optimal for φ, D, and $\vec{\alpha}$, and $P(\varphi, D, \vec{\alpha}, f, c, k)$ holds by assumption, this sequence S must contain $\geq ck$ test of v' and $\leq f(k)$ test of v.

All that remains to show that φ exhibits RI in the game $G(\varphi, D, k, \vec{\alpha}, g, b)$ is to show that all optimal strategies have positive expected payoff. To do this, it suffices to show that there is a strategy that has positive expected payoff. Let S be an arbitrary test-outcome sequence in $\mathcal{S}_k$. Without loss of generality, we can assume that $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > 1/2$. Let σ_S be the strategy that tests every variable the number of times that it occurs in S in the order that the variables occur in S, and guesses that the formula is true if and only if S was in fact the outcome sequence observed (and makes no guess otherwise). Since S will be observed with positive probability, it follows from Lemma 3.3 that σ_S has positive expected payoff. This completes the proof. □

It is easy to show that all that affects $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ is the number of number of times that each variable is tested and the outcome of the test, not the order in which the tests were made. It turns out that to determine whether a formula φ exhibits RI, we need to consider, for each truth assignment A that satisfies φ and test-outcome sequence S, the *A-trace* of S; this is a tuple that describes, for each variable v_i, the fraction of times v_i is tested and the outcome agrees with $A(v_i)$ compared to the fraction of times that the outcome disagrees with $A(v_i)$.

In Appendix B, we show that whether a formula exhibits RI can be determined by considering properties of the A-traces

of outcome sequences. Specifically, we show that the set of A-traces of optimal outcome sequences tends to a convex polytope as the length of S increases. This polytope has a characterisation as the solution set of an $O(n2^n)$-sized linear program (LP), so we can find points in the polytope in time polynomial in 2^n. Moreover, conditions such as a variable v is ignored while a variable v' that is no more relevant than v is not ignored correspond to further conditions on the LP, and thus can also be checked in time polynomial in 2^n. It follows that we can get a sufficient condition for a formula to exhibit RI or not exhibit RI by evaluating a number of LPs of this type.

Using these insights, we were able to exhaustively test all formulae that involve at most 4 variables to see whether, as the number of tests in the game increases, optimal strategies were testing a more relevant variable a negligible number of times relative to a less variable. Since the criterion that we use is only a sufficient condition, not a necessary one, we can give only a lower bound on the true number of formulae that exhibit RI. In the full paper, we discuss an additional conjecture, the *noise transfer conjecture* (NTC); if it holds, we can establish RI for significantly more formulae.

In the following table, we summarise our results. The first column lists the number of formulae that we are certain exhibit RI; the second column lists the number of additional formulae that exhibit RI if the NTC holds; the third column lists the remaining formulae, whose status is unknown. (Since RI is a semantic condition, when we say "formula", we really mean "equivalence class of logically equivalent formulae". There are 2^{2^n} equivalence classes of formulae with n variables, so the sum of the three columns in the row labeled n is 2^{2^n}.) As the results show, at least 15% of formulae exhibit RI, and this number increases to roughly 30% with the NTC.

n	exhibit RI	NTC $\Rightarrow$ RI	unknown
1	0	0	4
2	8	0	8
3	40	56	160
4	9952	8248	47334

Given the numbers involved, we could not exhaustively check what happens for $n \geq 5$. However, we did randomly sample 4000 formulae that involved n variables for $n = 5, \ldots, 9$. This is good enough for statistical reliability: we can model the process as a simple random sample of a binomially distributed parameter (the presence of RI), and in the worst case (if its probability in the population of formulae is exactly $\frac{1}{2}$), the 95% confidence interval still has width $\leq z\sqrt{\frac{1}{4000}\frac{1}{2}\left(1-\frac{1}{2}\right)} \approx 0.015$, which is well below the fractions of formulae exhibiting RI that we observe (all above 0.048). As the following table shows, RI continued to be quite common. Indeed, even for formulae with 9 variables, about 5% of the formulae we sampled exhibited RI.

n	exhibit RI	NTC $\Rightarrow$ RI	unknown
5	585	313	3102
6	506	138	3356
7	293	63	3644
8	234	30	3736
9	194	10	3796

The numbers suggest that the fraction of formulae exhibiting RI decreases as the number of variables increases. However, since the formulae that characterise situations of interest to people are likely to involve relatively few variables (or have a structure like disjunction or conjunction that we know exhibits RI), this suggests that RI is a widespread phenomenon. Indeed, if we weaken the notion of RI slightly (in what we believe is quite a natural way!), then RI is even more widespread. As noted in Example 4.5, formulae of the form $v_1 \vee (\neg v_1 \wedge v_2 \wedge \ldots \wedge v_n)$ do not exhibit RI in the sense of our definition. However, for these formulae, if we choose the payoffs b and g appropriately, an optimal strategy may start by testing v_1, but if sufficiently many test outcomes are $v_1 \approx F$, it will then try to establish that the formula is false by focussing on one variable of the conjunction $(v_2 \wedge \ldots \wedge v_n)$, and ignoring the rest. Thus, for all optimal strategies, we would have RI, not for all test-outcome sequences (i.e., not in all runs of the game), but on a set of test-outcome sequences that occur with positive probability.

We found it hard to find formulae that do not exhibit RI in this weaker sense. In fact, we conjecture that the only family of formulae that do not exhibit RI in this weaker sense are equivalent to XORs in zero or more variables $(v_1 \oplus \ldots \oplus v_n)$ and their negations (Note that this family of formulae includes v_i and $\neg v_i$.) If this conjecture is true, we would expect to quite often see rational agents (and decision-making computer programs) ignoring relevant variables in practice.

5 Testing as a measure of complexity

The notion of associating some "intrinsic difficulty" with concepts (typically characterised using Boolean formulae) has been a topic of continued interest in the cognitive science community [Vigo, 2011; Feldman, 2006; Love, Medin, and Gureckis, 2004; Shepard, Hovland, and Jenkins, 1961]. We can use our formalism to define a notion of difficulty for concepts. Our notion of difficulty is based on the number of tests that are needed to guarantee a positive expected payoff for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$. This will, in general, depend on D, $\vec{\alpha}$, g, and b. Actually, by Lemma 3.3, what matters is not g and b, but $q(b, g)$ (the threshold determined by g and b). Thus, our complexity measure takes D, $\vec{\alpha}$, and q as parameters.

Definition 5.1. Given a formula φ, accuracy vector $\vec{\alpha}$, distribution D, and threshold $0 < q \leq \frac{1}{2}$, the *$(D, q, \vec{\alpha})$-test complexity* $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi)$ of φ is the least k such that there exists a strategy with positive payoff for $G(\varphi, D, k, \vec{\alpha}, g, b)$, where g and b are chosen such that $q(b, g) = q$. □

To get a sense of how this definition works, consider what happens if we consider all formulae that use two variables, v_1 and v_2, with the same settings as in Example 2.1: $\vec{\alpha} = (1/4, 1/4)$, D is the uniform distribution on assignments, $g = 1$, and $b = -16$:

1. If φ is simply T or F, any strategy that guesses the appropriate truth value, regardless of test outcomes, is optimal and gets a positive expected payoff, even when $k = 0$.
2. If φ is a single-variable formula of the form v_1 or $\neg v_1$, then the greatest certainty $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2|$ that is

attainable with any sequence of two tests is $2/5$, when $S = (v_1 \approx T, v_1 \approx T)$ or the same with F. This is smaller than $q(b, g)$, and so it is always optimal to make no guess; that is, all strategies for the game with $k = 2$ have expected payoff at most 0. If $k = 3$ and $S = (v_1 \approx T, v_1 \approx T, v_1 \approx T)$, then $(\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2) = 13/28 > q(b, g)$. Thus, if $k = 3$, the strategy that test v_1 three times and guesses the appropriate truth value iff all three tests agree has positive expected payoff.

3. If φ is $v_1 \oplus v_2$, then the shortest test-outcome sequences S for which $\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2$ is greater than $q(b, g)$ have length 7, and involve both variables being tested. Hence, the smallest value of k for which strategies with payoff above 0 exist is 7.

It is not hard to see that T and F have complexity 0, while disjunctions, conjunctions, and, more generally, majority ("m out of n variables are true") have low complexity. We also completely characterise the most difficult concepts, according to our complexity measure, at least in the case of a uniform distribution D_u on truth assignments (which is the one most commonly considered in practice).

Theorem 5.2. *Among all Boolean formulae in n variables, for all $0 < q \leq \frac{1}{2}$ and accuracy vectors $\vec{\alpha}$, the $(D_u, q, \vec{\alpha})$-test complexity is maximised by formulae equivalent to the n-variable XOR $v_1 \oplus \ldots \oplus v_n$ or its negation.*

Proof sketch. Call a formula φ *antisymmetric* in variable v if $\varphi(A) = \neg\varphi(A')$ for all pairs of assignments A, A' that only differ in the truth value of v. It is easy to check that if a formula is antisymmetric in all variables, it is equivalent to an XOR or a negation of one. Given a formula φ, the *antisymmetrisation φ_v of φ along v* is the unique formula such that $\varphi_v(A) = \varphi(A)$ if $A(v) = T$ and $\varphi_v(A) = \neg\varphi(A[v \mapsto T])$ otherwise. We can show that the $(D_u, q, \vec{\alpha})$-test complexity of φ_v is at least as high as that of φ, and that if $v' \neq v$, then φ_v is antisymmetric in v' iff φ is antisymmetric in v'. So, starting with an arbitrary formula φ, we antisymmetrise every variable in turn. We then end up with an XOR or the negation of one. Moreover, each antisymmetrisation step in the process gives a formula whose test complexity is at least as high as that of the formula in the previous step. The desired result follows. A detailed proof can be found in Appendix C. □

It is of interest to compare our notion of "intrinsic difficulty" with those considered in the cognitive science literature. That literature can broadly be divided up into purely experimental approaches, typically focused on comparing the performance of human subjects in dealing with different categories, and more theoretical ones that posit some structural hypothesis regarding which categories are easy or difficult.

The work of Shepard, Hovland, and Jenkins [1961] is a good example of the former type; they compare concepts that can be defined using three variables in terms of how many examples (pairs of assignments and corresponding truth values of the formula) it takes human subjects to understand and remember a formula φ, as defined by a subject's ability to predict the truth value of φ correctly for a given truth assignment. We can think of this work as measuring how hard it is to work with a formula; our formalism is measuring how hard it is to learn the truth value of a formula. The difficulty ranking found experimentally by Shepard et al. mostly agrees with our ranking, except that they find two- and three-variable XORs to be easier that some other formulae, whereas we have shown that these are the hardest formulae. Perhaps this is suggesting that there are differences between how hard it is to work with a concept and how hard it is to learn it.

Feldman [2006] provides a good example of the latter approach. He proposes the notion of the *power spectrum* of a formula φ. Roughly speaking, this counts the number of antecedents in the conjuncts of a formula when it is written as a conjunction of implications where the antecedent is a conjunction of literals and the conclusion is a single literal. For example, the formula $\varphi = (v_1 \wedge (v_2 \vee v_3)) \vee (\neg v_1 \wedge (\neg v_2 \wedge \neg v_3))$ can be written as the conjunction of three such implications: $(v_2 \rightarrow v_1) \wedge (v_3 \rightarrow v_1) \wedge (\neg v_2 \wedge v_1 \rightarrow v_3)$. Since there are no conjuncts with 0 antecedents, 2 conjuncts with 1 antecedent, and 1 conjunct with 2 antecedents, the power spectrum of φ is $(0, 1, 2)$. Having more antecedents in an implication is viewed as making concepts more complicated, so a formula with a power spectrum of $(0, 1, 1)$ is considered more complicated than one with a power spectrum of $(0, 3, 0)$, and less complicated than one with a power spectrum of $(0, 0, 3)$.

A formula with a power spectrum of the form $(i, j, 0, \ldots, 0)$ (i.e., a formula that can be written as the conjunction of literals and formulae of the form $x \rightarrow y$, where x and y are literals) is called a *linear category*. Experimental evidence suggests that human subjects generally find linear categories easier to learn than nonlinear ones [Feldman, 2006; Love, Medin, and Gureckis, 2004]. (This may be related to the fact that such formulae are linearly separable, and hence learnable by support vector machines [Vapnik and Lerner, 1963].) Although our complexity measure does not completely agree with the notion of a power spectrum, both notions classify XORs and their negations as the most complex; these formulae can be shown to have a power spectrum of the form $(0, \ldots, 0, 2^{n-1})$.

Another notion of formula complexity is the notion of *subjective structural complexity* introduced by Vigo [2011], where the subjective structural complexity of a formula φ is $|Sat(\varphi)|e^{-\|\vec{f}\|_2}$, where $Sat(\varphi)$ is the set of truth assignments that satisfy φ, $f = (f_1, \ldots, f_n)$, f_i is the fraction of truth assignments that satisfy φ such that changing the truth value of v_i results in a truth assignment that does not satisfy φ, and $\|\vec{f}\|_2 = \sqrt{(f_1)^2 + \cdots + (f_n)^2}$ represents the ℓ^2 norm. Unlike ours, with this notion of complexity, φ and $\neg\varphi$ may have different complexity (because of the $|Sat(\varphi)|$ factor). However, as with our notion, XORs and their negation have maximal complexity.

6 Conclusion

We have presented the information-acquisition game, a game-theoretic model of gathering information to inform a decision whose outcome depends on the truth of a Boolean formula. We argued that it is hard to find optimal strategies for this model by brute force, and presented the random-test heuristic, a simple strategy that only has weak guarantees but is

computationally tractable. It is an open question whether better guarantees can be proven for the random-test heuristic, and whether better approaches to testing that are still more computationally efficient than brute force exist.

We used our techniques to show that RI is a widespread phenomenon (at least, for formulae that use at most 9 variables, which certainly covers most naturally-arising concepts for humans). We hope in future work to get a natural structural criterion for when formulae exhibit RI that can be applied to arbitrary formulae.

Finally, we discussed how the existence of good strategies in our game can be used as a measure of the complexity of a Boolean formula. It would be useful to get a better understanding of whether test complexity captures natural structural properties of concepts.

Although we have viewed the information-acquisition game as a single-agent game, there are natural extensions of it to multi-agent games, where agents are collaborating to learn about a formula. We could then examine different degrees of coordination for these agents. For example, they could share information at all times, or share information only at the end (before making a guess). The goal would be to understand whether there is some structure in formulae that makes them particularly amenable to division of labour, and to what extent it can be related to phenomena such as rational inattention (which may require the agents to coordinate on deciding which variable to ignore).

Acknowledgements

We thank David Goldberg, David Halpern, Bobby Kleinberg, Dana Ron, Sarah Tan, and Yuwen Wang as well as the anonymous reviewers for helpful feedback, discussions and advice. This work was supported in part by NSF grants IIS-1703846 and IIS-1718108, AFOSR grant FA9550-12-1-0040, ARO grant W911NF-17-1-0592, and a grant from the Open Philanthropy project.

References

Druzdzel, M. J., and Suermondt, H. J. 1994. Relevance in probabilistic models: "Backyards" in a "small world". In *Working notes of the AAAI–1994 Fall Symposium Series: Relevance*, 60–63.

Feldman, J. 2006. An algebra of human concept learning. *Journal of Mathematical Psychology* 50(4):339 – 368.

Lang, J.; Liberatore, P.; and Marquis, P. 2003. Propositional independence – formula-variable independence and forgetting. *Journal of Artificial Intelligence Research* 18:391–443.

Love, B. C.; Medin, D. L.; and Gureckis, T. M. 2004. Sustain: A network model of category learning. *Psychological Review* 111(2):309–332.

Shepard, R. N.; Hovland, C. I.; and Jenkins, H. M. 1961. Learning and memorization of classifications. *Psychological Monographs: General and Applied* 75(3):1–42.

Sims, C. A. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50(3):665–690.

Umans, C. 1999. On the complexity and inapproximability of shortest implicant problems. In *Proc. of Automata, Languages and Programming: 26th International Colloquium (ICALP '99)*, 687–696. Berlin, Heidelberg: Springer.

Vapnik, V. N., and Lerner, A. Y. 1963. Recognition of patterns using generalized portraits. *Avtomat. i Telemekh.* 24:774–780.

Vigo, R. 2011. Representational information: a new general notion and measure of information. *Information Sciences* 181:4847–4859.

Wiederholt, M. 2010. Rational inattention. In Blume, L., and Durlauf, S., eds., *The New Palgrave Dictionary of Economics (online edition)*. New York: Palgrave Macmillan.

A Calculations for Example 2.1

In this section, we fill in the details of the calculations for Example 2.1.

We have

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx \mathrm{T}) \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx \mathrm{T})}{\Pr_{D,\vec{\alpha}}(x \approx \mathrm{T})} \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx \mathrm{T})}{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx \mathrm{T}) + \Pr_{D,\vec{\alpha}}(x = \mathrm{F} \cap x \approx \mathrm{T})} \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(x \approx \mathrm{T} \mid x = \mathrm{T}) \Pr_{D,\vec{\alpha}}(x = \mathrm{T})}{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx \mathrm{T}) + \Pr_{D,\vec{\alpha}}(x = \mathrm{F} \cap x \approx \mathrm{T})} \\
= \; & \frac{(3/4)(1/2)}{(3/4)(1/2) + (1/4)(1/2)} \\
= \; & 3/4.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap x \approx_2 \mathrm{T}) \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx_2 \mathrm{T} \mid x \approx_1 \mathrm{T})}{\Pr_{D,\vec{\alpha}}(x \approx_2 \mathrm{T} \mid x \approx_1 \mathrm{T})} \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx_2 \mathrm{T} \mid x \approx_1 T)}{\Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx_2 \mathrm{T} \mid x \approx_1 \mathrm{T}) + \Pr_{D,\vec{\alpha}}(x = \mathrm{F} \cap x \approx_2 \mathrm{T} \mid x \approx_1 \mathrm{T})} \\
= \; & \frac{(3/4)(3/4)}{(3/4)(3/4) + (1/4)(1/4)} \\
= \; & 9/10,
\end{aligned}$$

where we use that

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \cap x \approx_2 \mathrm{T} \mid x \approx_1 \mathrm{T}) \\
= \; & \Pr_{D,\vec{\alpha}}(x \approx_2 \mathrm{T} \mid x = \mathrm{T}) \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx_1 \mathrm{T})
\end{aligned}$$

by conditional independence.

Hence

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(\varphi = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap x \approx_2 \mathrm{T}) \\
= \; & 1 - (1 - \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap x \approx_2 \mathrm{T})) \cdot (1 - \Pr_{D,\vec{\alpha}}(y = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap x \approx_2 \mathrm{T})) \\
= \; & 1 - (1 - \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap x \approx_2 \mathrm{T}))(1 - \Pr_{D,\vec{\alpha}}(y = \mathrm{T})) \\
= \; & 1 - (1/10)(1/2) \\
= \; & 19/20 \\
> \; & 16/17.
\end{aligned}$$

The same inequality holds if we observe that y is true twice. However,

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(\varphi = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap y \approx_2 \mathrm{T}) \\
= \; & 1 - (1 - \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap y \approx_2 \mathrm{T})) \cdot (1 - \Pr_{D,\vec{\alpha}}(y = \mathrm{T} \mid x \approx_1 \mathrm{T} \cap y \approx_2 \mathrm{T})) \\
= \; & 1 - (1 - \Pr_{D,\vec{\alpha}}(x = \mathrm{T} \mid x \approx \mathrm{T}))(1 - \Pr_{D,\vec{\alpha}}(y = \mathrm{T} \mid y \approx \mathrm{T})) \\
= \; & 1 - (1/4)(1/4) \\
= \; & 15/16 \\
< \; & 16/17,
\end{aligned}$$

as tests of each variable are independent of the truth value of all other variables. Clearly, if any test come out "false", the conditional probability that the formula is true will be even lower. So indeed, measuring the same variable twice is strictly better than measuring each of them once. Given the way we have set up the payoffs, in the latter case the only rational action is to play safe and not make a guess, making the expected payoff 0.

B Quantifying rational inattention

Our goal is to show that a large proportion of Boolean formulae exhibit RI. Proposition 4.7 gives a sufficient criterion in terms of the structure of the optimal sequences of test outcomes of each length. To make use of this criterion, we introduce some machinery to reason about optimal sequences of test outcomes. The key definition turns out to be that of the *characteristic fraction* of S for φ, denoted $\mathrm{cf}(\varphi, s)$, which is a quantity that is inversely ordered to $\Pr_{D,\vec{\alpha}}(\varphi|S)$ (Lemma B.4) (so the probability is maximised iff the characteristic fraction is minimised and vice versa), while exhibiting several convenient properties that enable the subsequent analysis. Let o_i represent the odds of making a correct observation of v_i, namely, the probability of observing $v_i \approx b$ conditional on v_i actually being b divided by the probability of observing $v_i \approx b$ condition on v_i not being b. If we assume that $o_i = o_j$ for all variables i and j, and let o represent this expression, then $\mathrm{cf}(\varphi, S)$ is the quotient of two polynomials, and has the the form

$$\frac{c_1 o^{d_1|S|} + \ldots + c_{2^n} o^{d_{2^n}|S|}}{e_1 o^{f_1|S|} + \ldots + e_{2^n} o^{f_{2^n}|S|}},$$

where c_j, d_j, e_j, and f_j are terms that depend on the truth assignment A_j, so we have one term for each of the 2^n truth assignments, and $0 \leq d_j, f_j \leq 1$. For a test-outcome sequence S that is optimal for φ, we can show that $f_j = 1$ for some j. Thus, the most significant term in the denominator (i.e., the one that is largest, for $|S|$ sufficiently large) has the form $co^{|S|}$. We call the factor d_i before $|S|$ in the exponent of the leading term of the numerator the *max-power* (Definition B.11) of the characteristic function. We can show that the max-power is actually independent of S (if S is optimal for φ). Since we are interested in the test-outcome sequence S for which $\mathrm{cf}(\varphi, S)$ is minimal (which is the test-outcome sequence for which $\Pr_{D,\vec{\alpha}}(\varphi|S)$ is maximal), for each k, we want to find that S of length k whose max-power is minimal. As we show, we can find the sequence S whose max-power is minimum by solving a linear program (Definition B.13).

Not only does each sequence S translate to an A-trace, but we can conversely also find (B.20), for any length k and any vector v that satisfies some sanity checks, a sequence $S_k(v)$ whose A-trace is approximately v (arbitrarily close to, as k gets large). Using this, we can formally state the necessary variant of the earlier statement that only leading terms matter when $|S|$ is large: if the A-trace of test-outcome sequence S is too far removed from any solution point of the LP, then so must be its max-power as a consequence of LP continuity (B.19), and hence (B.21) $|S|$ is either small or S is not optimal, as approximations $S_k(v)$ to any solution point v have a higher-order leading term and so in fact can be shown to give higher conditional probability of φ's truth or falsity.

So as optimal test sequences S get longer, the set of their A-traces converges to the set of solutions to the LP described above (in the sense that there is a negligible function δ bounding their distance from this set depending on the length $|S|$). It turns out that this lifts (B.16) the criterion of Proposition 4.7 to a condition on the LP solution set: when a fixed entry is zero in all LP solution points, then δ bounds above the number of times the variable corresponding to it can be measured in an optimal test-outcome sequence (and so the variable is necessarily ignored by any such sequence), and when another entry is $C > 0$, the corresponding variable may not be ignored. Therefore, the existence of such entries in all solution points to the LP is sufficient to conclude that all optimal test-outcome sequences for φ satisfy the precondition of 4.7, and hence φ exhibits rational inattention.

B.1 Computing $\Pr_{D,\vec{\alpha}}(A \mid S)$

In this subsection, we provide a representation of $\Pr_{D,\vec{\alpha}}(A \mid S)$ for an assignment A and a test-outcome sequence S. that makes the calculation of whether a formula exhibits RI easier. As the notation suggests, the lemma also shows that the probability is independent of the strategy σ. We use the following abbreviations:

- $o_i = \frac{1/2+\alpha_i}{1/2-\alpha_i}$. We can think of o_i as the odds of making a correct observation of v_i; namely, the probability of observing $v_i \approx b$ conditional on v_i actually being b divided by the probability of observing $v_i \approx b$ condition on v_i not being b.
- $n^+_{S,A,i} = |\{S[j] = (v_i \approx A(v_i))\}|$. Thus, $n^+_{S,A,i}$ is the number of times that v_i is observed to have the correct value according to truth assignment A in test-outcome sequence S.
- $r_{D,\vec{\alpha}}(A, S) = \Pr_{D,\vec{\alpha}}(A) \prod_{i=1}^{n} o_i^{n^+_{S,A,i}}$

We further adopt the convention of treating a formula φ as a function from truth assignments to truth values. A truth assignment itself, as the name suggests, is a function that assigns each variable a truth value. We can therefore write $A(v_1) = T$ and $A(v_2) = F$ to denote that the assignment A sets v_1 to be true and v_2 to be false, and $\varphi(A) = F$ for the formula $\varphi = v_1 \wedge v_2$ to signify that φ is false on the assignment A (i.e. when $v_1 = T$ and $v_2 = F$).

Lemma B.1. *For all accuracy vectors $\vec{\alpha}$, product distributions D, assignments A and test-outcome sequences S,*

$$\Pr_{D,\vec{\alpha}}(A \mid S) = \frac{r_{D,\vec{\alpha}}(A, S)}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S)}.$$

Thus,

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S) = \sum_{\{A:\varphi(A)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(A \mid S) = \frac{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S)}{\sum_{A'} r_{D,\vec{\alpha}}(A', S)}.$$

These probabilities do not depend on the strategy σ.

Proof. By Bayes' rule, for all truth assignments A and sequences $S = [v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k]$ of test outcomes, we have

$$\begin{aligned} \Pr_{D,\vec{\alpha},\sigma}(A \mid S) &= \frac{\Pr_{D,\vec{\alpha},\sigma}(S \mid A)\Pr_{D,\vec{\alpha}}(A)}{\Pr_{D,\vec{\alpha},\sigma}(S)} \\ &= \frac{\Pr_{D,\vec{\alpha},\sigma}(S \mid A)\Pr_{D,\vec{\alpha}}(A)}{\sum_{\text{truth assignments } A'} \Pr_{D,\vec{\alpha},\sigma}(S \mid A')\Pr_{D,\vec{\alpha}}(A')}. \end{aligned} \tag{2}$$

Suppose that $S = (v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k)$. For an arbitrary truth assignment A',

$$\begin{aligned} \Pr_{D,\vec{\alpha},\sigma}(S \mid A') = \; & \sigma(v_{i_1} \text{ chosen}) \Pr_{D,\vec{\alpha},\sigma}(v_{i_1} \approx b_1 \text{ observed} \mid v_{i_1} \text{ chosen}, A') \ldots \\ & \sigma(v_{i_k} \text{ chosen} \mid (v_{i_1} \approx b_{i_1}, \ldots, v_{i_{k-1}} \approx b_{i_{k-1}}) \text{ observed}) \\ & \Pr_{D,\vec{\alpha},\sigma}(v_{i_k} \approx b_{i_k} \text{ observed} \mid (v_{i_1} \approx b_{i_1}, \ldots, v_{i_{k-1}} \approx b_{i_{k-1}}) \text{ observed}, v_{i_k} \text{ chosen}, A') \end{aligned}$$

Notice that the terms $\sigma(v_{i_1} \text{ chosen}), \ldots, \sigma(v_{i_k} \text{ chosen} \mid (v_{i_1} \approx b_{i_1}, \ldots, v_{i_{k-1}} \approx b_{i_{k-1}} \text{ observed})$ are common to $\Pr_{D,\vec{\alpha},\sigma}(S \mid A')$ for all truth assignments A', so we can pull them out of the numerator and denominator in (2) and cancel them. Moreover, the probabilities $\Pr_{D,\vec{\alpha},\sigma}(v_{i_1} \approx b_1 \text{ observed} \mid v_{i_1} \text{ chosen}, A'), \ldots, \Pr_{D,\vec{\alpha},\sigma}(v_{i_k} \approx b_{i_k} \text{ observed} \mid (v_{i_1} \approx b_{i_1}, \ldots, v_{i_{k-1}} \approx b_{i_{k-1}}) \text{ observed}, v_{i_k} \text{ chosen}, A')$ do not depend on σ, so we can drop the σ from the subscript of $\Pr_{D,\vec{\alpha},\sigma}$, and are independent of what earlier observations. Thus, it follows that

$$\Pr_{D,\vec{\alpha},\sigma}(A \mid S) = \frac{\prod_{j=1}^{k} \Pr_{D,\vec{\alpha}}(v_{i_j} \approx b_j \text{ observed} \mid v_{i_j} \text{ chosen}, A)\Pr_{D,\vec{\alpha}}(A)}{\sum_{\text{truth assignments } A'} \prod_{j=1}^{k} \Pr_{D,\vec{\alpha}}(v_{i_j} \approx b_j \text{ observed} \mid v_{i_j} \text{ chosen}, A')\Pr_{D,\vec{\alpha}}(A')}.$$

Consider the numerator of this expression If $b_j = A(v_{i_j})$, then the jth term in the product is o_{i_j}; if $b_j = \neg A(v_{i_j})$, then the jth term in the product is 1. It easily follows that this expression is just $r_{D,\vec{\alpha}}(A, S)$. A similar argument shows that denominator is $\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S)$. This proves the first and third statements in the lemma. The second statement is immediate from the first. □

The next lemma gives some an intuitively natural property of those test-outcome sequences S that are *optimal* for φ, D, and $\vec{\alpha}$.

Lemma B.2. *If S is a test-outcome sequence that is optimal for φ, D, and $\vec{\alpha}$, and $\Pr_{D,\vec{\alpha}}(\varphi|S) \neq \Pr_{D,\vec{\alpha}}(\varphi) > 0$, then S does not contain observations both of the form $v_i \approx T$ and of the form $v_i \approx F$ for some v_i.*

Proof. Suppose that S is optimal for φ, D, and $\vec{\alpha}$, $\Pr_{D,\vec{\alpha}}(\varphi \mid S) \neq \Pr_{D,\vec{\alpha}}(\varphi)$, there are $n_1 > 0$ instance of $v_i \approx T$ in S, and $n_2 > 0$ instances of $v_i \approx F$ in S. Without loss of generality, suppose that $n_1 > n_2$. Suppose, by way of contradiction, that $n_2 > 0$. Let S_0 be the sequence that results from S by removing the n_2 occurrences of $v_i \approx F$ and the last n_2 occurrences of $v_i \approx T$. Thus, $|S_0| = |S| - 2n_2 < |S|$. It is easy to see that, for each truth assignment A, we have $n^+_{S,A,i} = n^+_{S_0,A,i} + n_2$. It thus follows from Lemma B.1 that $\Pr_{D,\vec{\alpha}}(\varphi \mid S) = \Pr_{D,\vec{\alpha}}(\varphi \mid S_0)$. We can similarly remove all other "contradictory" observations to get a sequence S_0 that does not contradict itself such that $|S_0| < |S|$ and $\Pr_{D,\vec{\alpha}}(\varphi \mid S) = \Pr_{D,\vec{\alpha}}(\varphi \mid S_0)$.

Suppose without loss of generality that $\Pr_{D,\vec{\alpha}}(\varphi) - 1/2 \geq 0$. Since it cannot be the case that for every test-outcome sequence S_0 of length $|S|$ we have $\Pr_{D,\vec{\alpha}}(\varphi \mid S_0) - 1/2 < \Pr_{D,\vec{\alpha}}(\varphi) - 1/2$, and S is optimal for φ, D, and $\vec{\alpha}$, we must have

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2 \geq |\Pr_{D,\vec{\alpha}}(\varphi) - 1/2|. \tag{3}$$

We want to show that we can add tests to S_0 to get a sequence S^* with $|S^*| = |S|$ such that $\Pr_{D,\vec{\alpha}}(\varphi \mid S^*) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_0) = \Pr_{D,\vec{\alpha}}(\varphi \mid S)$. This will show that S is not optimal for φ, D, and $\vec{\alpha}$, giving us the desired contradiction.

Suppose that $S_0 = (v_{i_1} \approx b_1, \ldots, v_k \approx b_k)$. Define test-outcome sequences $S_1, \ldots, S_k$ inductively by taking S_j to be S_{j-1} with $v_{i_j} \approx b_j$ removed if $\Pr_{D,\vec{\alpha}}(\varphi \mid S_{j-1}) \leq \Pr_{D,\vec{\alpha}}(\varphi \mid S_{j-1} \backslash (v_{i_j} \approx b_j))$ and otherwise taking $S_j = S_{j-1}$. It is immediate from the construction that $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k) \geq \Pr_{D,\vec{\alpha}}(\varphi \mid S_0) = \Pr_{D,\vec{\alpha}}(\varphi \mid S)$ and $|S_k| \leq |S_0| < |S|$. It cannot be the case that $|S_k| = 0$, for then $\Pr_{D,\vec{\alpha}}(\varphi) \geq \Pr_{D,\vec{\alpha}}(\varphi \mid S)$. Since $\Pr_{D,\vec{\alpha}}(\varphi) \neq \Pr_{D,\vec{\alpha}}(\varphi \mid S)$ by assumption, we would have $\Pr_{D,\vec{\alpha}}(\varphi) > \Pr_{D,\vec{\alpha}}(\varphi \mid S)$, contradicting (3).

Suppose that $v_i \approx b$ is the last test in S_k. Let $S_k^- = S_k \backslash (v_i \approx b)$, so that $S_k = S_k^- \cdot (v_i \approx b)$. By construction, $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k^-)$. That is, observing $v \approx b$ increased the conditional probability of φ. We now show that observing $v \approx b$ more often increases increases the conditional probability of φ further; that is, for all m, $\Pr_{D,\vec{\alpha}}(\varphi \mid (S_k \cdot (v_i \approx b)^m) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k)$. We can thus take $S^* = (S_k \cdot (v_i \approx b)^{|S|-|S_k|})$.

It follows from Lemma B.1 that

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S_k) = \sum_{\{A:\varphi(A)=T\}} \Pr_{D,\vec{\alpha}}(A \mid S_k) = \frac{\sum_{\{A:\varphi(A)=T\}} r_{D,\vec{\alpha}}(A, S_k)}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S_k)}$$

$$\text{and } \Pr_{D,\vec{\alpha}}(\varphi \mid S_k^-) = \sum_{\{A:\varphi(A)=T\}} \Pr_{D,\vec{\alpha}}(A \mid S_k^-) \frac{\sum_{\{A:\varphi(A)=T\}} r_{D,\vec{\alpha}}(A, S_k^-)}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S_k^-)}.$$

Note that for a truth assignment A', $r_{D,\vec{\alpha}}(A', S_k^-) = r_{D,\vec{\alpha}}(A', S_k)$ if $A(v_i) \neq b$ and A', $r_{D,\vec{\alpha}}(A', S_k^-) = o_i r_{D,\vec{\alpha}}(A', S_k)$ if $A(v_i) = b$. Thus, there exist x_1, x_2, y_1, y_2 such that $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k^-) = \frac{x_1+x_2}{y_1+y_2}$ and $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k) = \frac{o_i x_1+x_2}{o_i y_1+y_2}$. Indeed, $x_1 = \sum_{\{A:\varphi(A)=T, A(v_i)=b\}} r_{D,\vec{\alpha}}(S_k^0, A)$, $x_2 = \sum_{\{A:\varphi(A)=T, A(v_i)\neq b\}} r_{D,\vec{\alpha}}(S_k^0, A)$, $y_1 = \sum_{\{A:A(v_i)=b\}} r_{D,\vec{\alpha}}(S_k^0, A)$, and $y_2 = \sum_{\{A:A(v_i)\neq b\}} r_{D,\vec{\alpha}}(S_k^0, A)$. Since $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k^-)$, we must have

$$\frac{o_i x_1 + x_2}{o_i y_1 + y_2} > \frac{x_1 + x_2}{y_1 + y_2}. \tag{4}$$

Since $x_1, x_2, y_1, y_2 \geq 0$, crossmultiplying shows that (4) holds iff

$$x_2 y_1 + o_i x_1 y_2 > x_1 y_2 + o_i x_2 y_1.$$

Similar manipulations show that

$$\begin{array}{ll} & \Pr_{D,\vec{\alpha}}(\varphi \mid S_k \cdot (v_i \approx b) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k) \\ \text{iff} & \dfrac{o_i^2 x_1 + x_2}{o_i 2 y_1 + y_2} > \dfrac{o_i x_1 + x_2}{o_i y_1 + y_2} \\ \text{iff} & x_2 y_1 + o_i x_1 y_2 > x_1 y_2 + o_i x_2 y_1. \end{array}$$

Thus, $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k \cdot (v_i \approx b)) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k)$. A straightforward induction shows that $\Pr_{D,\vec{\alpha}}(\varphi \mid S_k \cdot (v_i \approx b)^h) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k)$ for all h, so $\Pr_{D,\vec{\alpha}}(\varphi \mid S^*) > \Pr_{D,\vec{\alpha}}(\varphi \mid S_k) = \Pr_{D,\vec{\alpha}}(\varphi \mid S)$, as desired. □

B.2 Characteristic fractions and the limit of traces

Definition B.3. The *characteristic fraction* of a test-outcome sequence S for φ is

$$\mathrm{cf}(\varphi, S) = \frac{\sum_{\{A:\varphi(A)=\mathrm{F}\}} r_{D,\vec{\alpha}}(A, S)}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S)}.$$

□

The importance of this quantity is due to the following:

Lemma B.4. $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > \Pr_{D,\vec{\alpha}}(\varphi \mid S')$ *iff* $\mathrm{cf}(\varphi, S) < \mathrm{cf}(\varphi, S')$.

Proof. Since $x < y \Leftrightarrow (1/x) > (1/y)$, it follows from Lemma B.1 that $\Pr_{D,\vec{\alpha}}(\varphi \mid M) < \Pr_{D,\vec{\alpha}}(\varphi \mid M')$ if and only if

$$\frac{\sum_A r_{D,\vec{\alpha}}(A, S)}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S)} > \frac{\sum_A r_{D,\vec{\alpha}}(A, S'))}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S')\}},$$

which is true if and only if

$$\frac{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S) + \sum_{\{A:\varphi(A)=\mathrm{F}\}} r_{D,\vec{\alpha}}(A, S)}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S)} > \frac{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S') + \sum_{\{A:\varphi(A)=\mathrm{F}\}} r_{D,\vec{\alpha}}(A, S')}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S')},$$

that is, if and only if

$$\frac{\sum_{\{A:\varphi(A)=\mathrm{F}\}} r_{D,\vec{\alpha}}(A, S)}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S)} > \frac{\sum_{\{A:\varphi(A)=\mathrm{F}\}} r_{D,\vec{\alpha}}(A, S')}{\sum_{\{A:\varphi(A)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S')}.$$

The statement of the lemma follows. □

Example B.5. Let $\varphi = (v_1 \wedge v_2) \vee (\neg v_2 \wedge \neg v_3)$ and $S = (v_2 \approx F, v_1 \approx T)$, and suppose that the prior D is uniform, so $o_1 = \ldots = o_n = o$. Then

$$\mathrm{cf}(\varphi, S) = \frac{o^1 + o^2 + o^0 + o^0}{o^1 + o^2 + o^1 + o^1}.$$

More generally, suppose that $S = ((v_1 \approx T)^{c_1 k}, (v_2 \approx F)^{c_2 k}, (v_3 \approx F)^{c_3 k})$ for some positive integer k and real constants $0 \leq c_1, c_2, c_3 \leq 1$ with $c_1 + c_2 + c_3 = 1$. Then

$$\mathrm{cf}(\varphi, S) = \frac{o^{c_2 k + c_1 k} + o^{c_1 k} + o^{c_3 k} + o^0}{o^{c_1 k} + o^{c_1 k + c_3 k} + o^{c_2 k + c_3 k} + o^{c_1 k + c_2 k + c_3 k}}.$$

□

Definition B.6. Given a test-outcome sequence S and truth assignment A, the *A-trace* of S, denoted $\mathrm{Tr}_A(S)$, is the n-dimensional vector $\mathrm{Tr}_A(S) = (n^+_{S,A,1}/|S|, \ldots, n^+_{S,A,n})$. □

Example B.7. Consider the sequence of test outcomes

$$S = (v_1 \approx T, v_2 \approx T, v_1 \approx T, v_1 \approx T, v_1 \approx F).$$

This sequence has 3 instances of $v_1 \approx T$, 1 instance of $v_1 \approx F$ and 1 instance of $v_2 \approx T$. So the $\{v_1 \mapsto T, v_2 \mapsto T\}$-trace of S is $(\frac{2}{5}, \frac{1}{5})$; the $\{v_1 \mapsto F, v_2 \mapsto T\}$-trace of S is $(-\frac{2}{5}, \frac{1}{5})$. The sequence

$$S' = [v_1 \approx T, v_2 \approx F, v_1 \approx T, v_1 \approx T, v_1 \approx T]$$

has 4 instances of $v_1 \approx T$ and 1 of $v_2 \approx F$, so the $\{v_1 \mapsto T, v_2 \mapsto F\}$-trace of S' is $(\frac{4}{5}, \frac{1}{5})$. □

Definition B.8. If $\vec{c} = (c_1, \ldots, c_n)$, φ is a formula in the n variables $v_1, \ldots, v_n$ and A is a truth assignment, then the *characteristic fraction of the A-trace* is the function cf_A, where

$$\mathrm{cf}_A(\varphi, \vec{c}, k) \quad = \quad \frac{\sum_{\{B:\varphi(B)=\mathrm{F}\}} \Pr_{D,\vec{\alpha}}(B) o_i^{\sum_{\{v_i:A(v_i)=B(v_i)\}} c_i k}}{\sum_{\{B:\varphi(B)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(B) o_i^{\sum_{\{v_i:A(v_i)=B(v_i)\}} c_i k}}.$$

□

Definition B.9. The test-outcome sequence S is *compatible with* truth assignment A if all test outcomes in S are consistent with A: that is, S contains *no* observations of the form $v_i \approx \neg A(v_i)$. □

$\mathrm{cf}(\varphi, S)$ and $\mathrm{cf}_A(\varphi, \vec{c}, k)$ are clearly closely related. The following lemma makes this precise.

Lemma B.10. *For all truth assignments A compatible with S, we have*

$$\mathrm{cf}(\varphi, S) = \mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|).$$

Proof. If A is compatible with S, then $(\mathrm{Tr}_A(S))_i = n^+_{S,A,i}/|S|$ for all i, so the result is immediate from the definition. □

Recall that our goal is to find optimal test-outcome sequences for φ, that is, sequences S that maximise $\Pr_{D,\vec{\alpha}}(\varphi|S)$. By Lemma B.4, this means that we want to minimise $\mathrm{cf}(\varphi, S)$. Since, by Lemma B.10, $\mathrm{cf}(\varphi, S) = \mathrm{cf}_A(\varphi, S)$ for all truth assignments A, it suffices to find a sequence S and a truth assignment A that is comptible with S for which $\mathrm{cf}_A(\varphi, S)$ is minimal.

Assumption: We assume for ease of exposition in the remainder of the paper that the measurement accurancy of each variable is the same, that is, $\alpha_1 = \cdots = \alpha_n$. This implies that $o_1 = \cdots = o_n$; we use o to denote this common value. While we do not need this assumption for our results, allowing non-uniform measurement vectors $\vec{\alpha}$ would lead to notions different notions of RI; the formulae that exhibit $\vec{\alpha} = (0.1, 0.1)$-RI might not be the same as those that exhibit $\vec{\alpha} = (0.1, 0.3)$-RI. With this assumption, we can show that $\mathrm{cf}_A(\varphi, S)$ is essentially characterised by the terms in its numerator and denominator with the largest exponents. Every optimal test-outcome sequence S is compatible with some assignment A. Since all test outcomes in S are consistent with A, if $\varphi(A) = T$, the summand due to A in the denominator of $\mathrm{cf}(\varphi, S) = \mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|)$ is of the form $\Pr_{D,\vec{\alpha}}(A) o^{|S|}$. This term must be the highest power of o that occurs in the denominator. The highest power of o in the numerator of $\mathrm{cf}_A(S)$ will in general be smaller than $1 \cdot |S|$, and depends on the structure of φ. As Lemma B.12 below shows, that power is characterised by the following function:

Definition B.11. The *max-power* of a vector $\vec{c} \in \mathbb{R}^n$ is

$$\mathrm{maxp}_{\varphi,A}(\vec{c}) = \max_{\{B:\varphi(B)\neq\varphi(A)\}} \sum_{\{i:A(v_i)=B(v_i)\}} c_i.$$

□

Lemma B.12. *If S is a test-outcome sequence compatible with A and $\varphi(A) = T$ (resp., $\varphi(A) = F$), then the highest power of o that occurs in the numerator of* $\mathrm{cf}(\varphi, S)$ *(resp.,* $\mathrm{cf}(\neg\varphi, S)$ *is* $|S|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))$.

Proof. This follows from the definition of $\mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|)$, the observation that if S is compatible with A then all entries in $\mathrm{Tr}_A(S)$ are non-negative, and Lemma B.10. □

We now show that max-power can be characterized using a linear program (LP). Note that if R is a finite subset of $\mathbb{R}$, then

$$\max R = \min\{m \mid \forall r \in R : r \leq m\} \tag{5}$$

Hence, finding the $\vec{c}$ with $\sum_i c_i = 1$ and $c_i \geq 0$ that maximises $\max_{\{B:\varphi(B)\neq\varphi(A)\}} \sum_{\{i:A(v_i)=B(v_i)\}} c_i$ is equivalent to finding the $\vec{c}$ and m that minimise m subject to $\max_{\{B:\varphi(B)\neq\varphi(A)\}} \sum_{\{i:A(v_i)=B(v_i)\}} c_i \leq m$, $\sum_i c_i = 1$, and $c_i \geq 0$ for all i. These latter constraints are captured by the following LP. In the description of the LP (and throughout the remainder of this section), we assume that φ is a formula that mentions n variables, and that these are variables are $v_1, \ldots, v_n$. We further assume that all truth assignments are truth assignments to these n variables.

Definition B.13. Given a formula φ and truth assignment A,
define the *conflict LP* $L_A(\varphi)$ to be the linear program

$$\begin{aligned}
&\text{minimise } m \\
&\text{subject to } \sum_{\{i|A(v_i)=B(v_i)\}} c_i \leq m \text{ for all } B \text{ such that } \varphi(B) \neq \varphi(A) \\
&\quad \sum_i c_i = 1. \\
&\quad c_i \geq 0 \text{ for } i = 1, \ldots, n; \\
&\quad 0 \leq m \leq 1.
\end{aligned}$$

□

The constraint $0 \leq m \leq 1$ is not necessary; since the c_i's are non-negative and $\sum_i c_i = 1$, the m that satisfies the constraints must be between 0 and 1. However, adding this constraint ensures that the set of tuples $(c_1, \ldots, c_n, m)$ that satisfy the constraints form a compact set.

We call this LP the *conflict* LP because we are considering assignments B that *conflict* with A, in the sense that φ takes a different truth value on them than it does on A. To reason about conflict LPs, we first introduce some notation.

Definition B.14. Suppose that L is a linear program in n variables minimising an objective function $f : \mathbb{R}^n \to \mathbb{R}$ subject to some constraints.

- The *feasible set* of L, *Feas*$(L) \subseteq \mathbb{R}^n$, is the set of points that satisfy all the constraints of L.
- The *minimum* of the LP, $\mathrm{MIN}(L)$, is the infimum $\inf_{p \in \textit{Feas}(L)} f(p)$ attained by the objective function over the points in *Feas*(L).
- The *solution polytope* of L, $\mathrm{OPT}(L) \subseteq$ *Feas*$(L) \subseteq \mathbb{R}^n$, is the set of all feasible points at which f attains the minimum, that is, $\mathrm{OPT}(L) = \{p \in \textit{Feas}(L) : f(p) = \mathrm{MIN}(L)\}$.

□

Our goal is to show that, with some appropriate tweaks, the solutions to the conflict LPs tell us enough about the structure of optimal test-outcome sequences to derive a sufficient condition for a formula to exhibit RI. The directed max-power $\mathrm{maxp}_{\varphi,A}$ captures constraints of the LP $L_A(\varphi)$; specifically, the vector $(\vec{c}, m)$ is in the feasible set of $L_A(\varphi)$ if and only if $m \geq \mathrm{maxp}_{\varphi,A}(\vec{c})$ and the entries of $\vec{c}$ sum to 1.

Roughly speaking, $\mathrm{MIN}(L_A(\varphi))$ tells us how well any sequence of test outcomes compatible with the assignment A can do. Since every optimal sequence is compatible with some assignment, we therefore can find the max-power of optimal sequences by considering the minimum of the minima of all LPs:

Definition B.15. For a formula φ, define the *minimax power* $\mathrm{MIN}^*(\varphi)$ to be the minimum of minima:

$$\mathrm{MIN}^*(\varphi) = \min_{\text{assignments } A} \mathrm{MIN}(L_A(\varphi)).$$

An assignment A, and the LP $L_A(\varphi)$, are *relevant* if $\mathrm{MIN}(L_A(\varphi)) = \mathrm{MIN}^*(\varphi)$. □

The significance of this quantity is formalised by the following theorem.

Theorem B.16. *If there exists a constant C such that for all relevant truth assignments A and all solution points $\vec{c} = (c_1, \ldots, c_n, m) \in \mathrm{OPT}(L_A(\varphi))$, there exist indices i and j such that $v_i \leq_\varphi v_j$, $c_i > C$, and $c_j = 0$, then φ exhibits RI.*

To prove Theorem B.16, we will show that the antecedent of the theorem implies the antecedent of Proposition 4.7. The next lemma is a first step towards this goal. Proposition 4.7 involves a condition on test sequences that intuitively says that some variable is tested often, but another variable that is at least as important is tested very little. This condition arises repeatedly in the following proof, so we attach a name to it.

Definition B.17. Given a constant c and negligible function f, a test-outcome sequence S is (f, c)-*good* if there exist variables v_i and v_j such that $v_i \geq_\varphi v_j$, S contains $\geq c|S|$ tests of v_j and $\leq f(|S|)$ tests of v_i. S is (f, c)-*bad* if it is not (f, c)-good. □

Using this notation, Proposition 4.7 says that a formula φ exhibits RI if Proposition 4.7 says that a formula φ exhibits RI if for all open-minded product distributions D and accuracy vectors $\vec{\alpha}$, there exists a negligible function f and $c > 0$ such that all test-outcome sequences optimal for φ, D, and $\vec{\alpha}$ are (f, c)-good. The contrapositive of Proposition 4.7 says that if a formula does not exhibit RI, then for for all f and c, there is an (f, c)-bad test-outcome sequence optimal for φ, D, and $\vec{\alpha}$. Bad sequences are counterexamples to RI. The next lemma allows us to "boost" such counterexamples if they exist: whenever we have a single bad sequence, we in fact have an infinite family of arbitarily long bad sequences that can be considered refinements of the same counterexample.

Lemma B.18. *If, for all negligible functions f and constants $c > 0$, there exists an (f, c)-bad test-outcome sequence that is optimal for φ, D, and $\vec{\alpha}$, then for all f and c, there exists an infinite sequence $\{S_k\}$ of (f, c)-bad optimal test-outcome sequences of increasing length (so that $|S_{k+1}| > |S_k|$), all optimal for φ, D, and $\vec{\alpha}$.*

Proof. We begin by proving the existence of an infinite sequence $\{S_k\}$ that satisfies all properties except possibly for the existence of the limit. To do this, we show the contrapositive. Fix φ, D, and $\vec{\alpha}$. We show that if there exist f and c for which there is no infinite sequence $\{S_k\}$ of (f, c)-bad test-outcome sequences optimal for φ, D, and $\vec{\alpha}$, then, for all D and $\vec{\alpha}$, there exist f'' and c'' for which there is not even a single (f'', c'')-bad test-outcome sequence that is optimal for φ, D, and $\vec{\alpha}$.

Choose f and c such that the premises of the contrapositive hold. Let $\mathcal{S}_{f,c}$ be the set of all (f, c)-bad test-outcome sequences that are optimal for φ, D and $\vec{\alpha}$. We can assume $\mathcal{S}_{f,c}$ is nonempty; otherwise we are clearly done. If there exist arbitrarily long sequences $S \in \mathcal{S}_{f,c}$, then we can pick a sequence $\{S_k\}$ of test-outcome sequences in $\mathcal{S}_{f,c}$ of increasing length from them, contradicting the assumption. In fact, this must be the case. For suppose, for the sake of contradiction, that it isn't. Then there must be an upper bound $\hat{k}$ on the lengths of test-outcome sequences in $\mathcal{S}_{f,c}$. Moreover, since there are only finitely many test-outcome sequences of a given length, $\mathcal{S}_{f,c}$ itself must also be finite. Thus,

$$c' = \min_{S \in \mathcal{S}_{f,c}} \max_{\text{variables } v_i \text{ in } \varphi} |(\mathrm{Tr}_A(S))_i|$$

is finite and greater than zero (as every sequence must test at least one variable and not contradict itself, so we are taking the minimum over finitely many terms greater than zero). Hence, $c'' = \min\{c, c'\}$ is also greater than 0. Let

$$f''(k) = \begin{cases} k & \text{if } k \leq \hat{k} \\ f(k) & \text{otherwise.} \end{cases}$$

Since f is negligible and f'' agrees with f for all $k > \hat{k}$, f'' is also negligible.

We claim that no test-outcome sequence S optimal for φ, D, and $\vec{\alpha}$ is (f'', c'')-bad. Indeed, all candidate sequences of length $|S| \leq \hat{k}$ are ruled out, because setting both v_i and v_j to be whatever variable is tested the most in S discharges the existential quantification of Definition B.17 (note $\leq_\varphi$ is a partial order, so $v_i \leq_\varphi v_i$ for all v_i) as the number of tests is bounded below by the minimum $c'|S|$ and above by the length $|S|$. Any test-outcome sequence S of length $|S| > \hat{k}$ must also be (f'', c'')-good. By choice of $\hat{k}$, S is (f, c)-good. Therefore, there must be a variable pair $v_i \geq_\varphi v_j$ such that S contains $\geq c|S|$ tests of v_j and $\leq f(|S|)$ tests of v_i. But $c'' \leq c$ by definition and $f''(|S|) = f(|S|)$, so v_i and v_j also bear witness to S being (f'', c'')-good. This gives the desired contradiction.

Thus, we have shown that there exists a sequence $\{S_k\}$ of bad test-sequence outcomes in $\mathcal{S}_{f,c}$ of increasing length. □

In the following, we use the standard notion of *1-norm*, where the 1-norm of a real-valued vector $\vec{v} = (v_1, \ldots, v_n)$ is

$$\|\vec{v}\|_1 = \sum_{i=1}^{n} |\vec{v}_i|,$$

the sum of absolute values of the entries of $\vec{v}$. We often consider the 1-norm of the difference of two vectors. Although the difference of vectors is defined only if they have same length, we occasionally abuse notation and write $\|\vec{v} - \vec{w}\|_1$ even when $\vec{v}$ and $\vec{w}$ are vectors of different lengths. In that case, we consider only the common components of the vectors. For example, if $\vec{v} = (v_1, \ldots, v_n)$ and $\vec{w} = (w_1, \ldots, w_m)$, then

$$\|\vec{v} - \vec{w}\|_1 = |v_1 - w_1| + \cdots + |v_{\min\{n,m\}} - w_{\min\{n,m\}}|.$$

The following fact about LPs will prove useful.

Lemma B.19. *If L is an LP with objective function f such that Feas(L) is compact, then for all $\epsilon > 0$, there exists an $\epsilon' > 0$ such that all feasible points $\vec{p} \in$ Feas(L) are either within ϵ of a solution point, that is,*

$$\exists \vec{o} \in \mathrm{OPT}(L).\, \|\vec{p} - \vec{o}\|_1 < \epsilon,$$

the objective function takes a value on them that is more than ϵ' away from the optimum,

$$|f(\vec{p}) - \mathrm{MIN}(L)| > \epsilon'.$$

Proof. Suppose not. Let Q be the set of all points in *Feas*(L) that do not satisfy the first inequality; that is,

$$Q = \{\vec{p} \in \textit{Feas}(L) \mid \forall \vec{o} \in \mathrm{OPT}(L).\|\vec{p} - \vec{o}\|_1 \geq \epsilon\}.$$

This set is bounded and closed, hence compact. If $\inf_{\vec{q} \in Q} |f(\vec{q}) - \mathrm{MIN}(L)| > 0$, then we can take $\epsilon' = \inf_{\vec{q} \in Q} |f(\vec{q}) - \mathrm{MIN}(L)| = 0$ since, for every point $\vec{P} \in PL(L)$, if $|(f(\vec{p}) = \mathrm{MIN}(L)| \leq \epsilon'$, then $\vec{p} \notin Q$, so $\vec{p}$ is within ϵ of some solution point.

So suppose that $\inf_{\vec{q}\in Q} |f(\vec{q}) - \mathrm{MIN}(L)| > 0$. Then there exists a sequence $(\vec{q}_i)_{i=1}^{\infty}$ of points in Q such that $\lim_{i\to\infty} f(\vec{q}) = \mathrm{MIN}(L)$. By the Bolzano-Weierstrass Theorem, this sequence must have a convergent subsequence $(\vec{q}'_i)_{i=1}^{\infty}$. Suppose that $\lim_{i\to\infty} \vec{q}'_i = \vec{q}^*$. This limit point is still in Q, as Q is compact. Since f is linear, hence continuous,

$$\begin{aligned} f(\vec{q}^*) &= f(\lim_{i\to\infty} \vec{q}'_i) \\ &= \lim_{i\to\infty} f(\vec{q}'_i) \\ &= \lim_{i\to\infty} f(\vec{q}_i) \\ &= \mathrm{MIN}(L). \end{aligned}$$

Thus, $\vec{q}^* \in \mathrm{OPT}(L)$ and $\vec{q}^* \in Q$, which is incompatible with the definition of Q. This gives the desired contradition. □

We have seen how to distill the information in a test-outcome sequence for a formula in n variables into an n-dimensional real vector by taking A-traces. The following lemma is to be understood as an approximate converse of this process: given an n-dimensional real vector, we construct a test-outcome sequence of a given length k whose A-trace is close (within an error term of $2n/k$) to that vector.

Lemma B.20. *If A is an assignment to the n variables of φ and $\vec{d} \in \mathbb{R}^n$ is such that all coordinates are non-negative and sum to 1, then for all $k \in \mathbb{N}$, there exists a test-outcome sequence $S_{k,\vec{d},A}$ of length k compatible with A such that $|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S_{k,\vec{d},A}) - \mathrm{maxp}_{\varphi,A}(\vec{d})| < 2n/k$.*

Proof. Define

$$S_{k,\vec{d},A} = ((v_1 \approx A(v_1))^{\lfloor d_1 k\rfloor}, \ldots, (v_n \approx A(v_n))^{\lfloor d_n k\rfloor}, (v_n \approx A(v_n))^{e}),$$

where $\lfloor x\rfloor$ is the floor of x (i.e., the largest integer n such that $n \le x$) and $e = k - (\sum_{i-1}^{n} \lfloor d_i k\rfloor)$ is whatever is needed to pad the sequence to having length k. (e.g., if $\vec{d} = (0.3, 0.7, \mathrm{MIN}^*(\varphi))$ and $k = 2$, then although the d_is sum to 1, $\lfloor d_1 k\rfloor = 0$ and $\lfloor d_2 k\rfloor = 1$, so we would have $e = 1$.)

Since $\sum_i d_i k = k$, $\sum_i \lfloor d_i k\rfloor \le k$, and hence $e \ge 0$. Also, $\mathrm{Tr}_A(S_{k,\vec{d},A})$ differs from $\vec{d}$ by at most $1/k$ in the first $n-1$ coordinates (as $|d_1 k - \lfloor d_1 k\rfloor| \le 1$) and by at most n/k in the final coordinate (as $e \le n$). Since, for each assignment B,

$$|\sum_{\{i:A(v_i)=B(v_i)\}} d_i - \sum_{\{i:A(v_i)=B(v_i)\}} (Tr_A(S_{k,\vec{d},A}))_i| \le (n-1)\frac{1}{k} + \frac{n}{k} \le \frac{2n}{k},$$

and for an arbitrary vector $\vec{c}$,

$$\mathrm{maxp}_{\varphi,A}(\vec{c}) = \max_{\{B:\varphi(B)\neq\varphi(A)\}} \sum_{\{i:A(v_i)=B(v_i)\}} c_i,$$

it follows that $|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S_{k,\vec{d},A}) - \mathrm{maxp}_{\varphi,A}(\vec{d})| < 2n/k$, as desired. □

We can finally relate the solutions of the conflict LP $L_A(\varphi)$ to traces to the traces of optimal test-outcome sequences. While the traces of optimal sequences may not be in $\mathrm{OPT}(L_A)$, they must get arbitrarily close to it as the length of the sequence gets larger.

Lemma B.21. *If D is open-minded,then there exists a function $\delta : \mathbb{N} \to \mathbb{R}$, depending only on φ, D, and $\vec{\alpha}$, such that*

- $\lim_{k\to\infty} \delta(k) = 0$ *and*
- *for all assignments A and test-outcome sequences S compatible with A that are optimal for φ, D, and $\vec{\alpha}$, the A-trace of S is within $\delta(|S|)$ of some solution $\vec{d} \in \mathrm{OPT}(L_A(\varphi))$, that is,*

$$\exists \vec{d} \in OPT(L_A(\varphi)) : \|\vec{d} - \mathrm{Tr}_A(S)\|_1 < \delta(|S|).$$

Proof. Fix φ, D, and $\vec{\alpha}$. Given $\epsilon > 0$, we show that there exists a constant k_ϵ such that for all truth assignments A and all test-outcome sequences S compatible with A such that $|S| > k_\epsilon$ and

$$\forall \vec{d} \in \mathrm{OPT}(L_A(\varphi)) : \|\mathrm{Tr}_A(S) - \vec{d}\|_1 \ge \epsilon \tag{6}$$

are not optimal for φ, D, and $\vec{\alpha}$. This suffices to prove the result, since we can then choose any descending sequence $\epsilon_0, \epsilon_1, \ldots$ and define $\delta(n) = \epsilon_n$ for all $k_{\epsilon_{n-1}} < n \le k_{\epsilon_n}$.

Fix $\epsilon > 0$ and A. Choose an arbitrary test-outcome sequence S compatible with A satisfying (6). Without loss of generality, we can assume that $\varphi(A) = T$. (Otherwise, run the proof for $\neg\varphi$ instead: every sequence that is optimal for one is optimal for the other, and $\neg\varphi(A) = T$.) Since the feasible set of the LP $L_A(\varphi)$ is compact by construction, by Lemma B.19, there

exists some $\epsilon_A > 0$ such that for all feasible points $p = (c_1, \ldots, c_n, m) \in \mathit{Feas}(L_A \cup \{m \leq 2\})$, either $\|p - \vec{d}\|_1 < \epsilon$ for some $\vec{d} \in \mathrm{OPT}(L)$, or $|m - \mathrm{MIN}(L_A)| > \epsilon_A$. Set

$$k_{\epsilon,A} = \max\left(\frac{4n}{\epsilon_A}, \frac{2}{\epsilon_A}\log_o \frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B)}\right).$$

(Since D is open-minded, $\min_B \Pr_{D,\vec{\alpha}}(B) > 0$, so this is well defined.) We now show that if $|S| > k_{\epsilon,A}$, then S is not optimal for φ, D, and $\vec{\alpha}$. We can then take $k_\epsilon = \max_A k_{\epsilon,A}$ to complete the proof.

Since S satisfies (6) by assumption, $\|(\mathrm{Tr}_A(S), \mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))) - \vec{d}\|_1 > \epsilon$ for all $\vec{d} \in \mathrm{OPT}(L_A)$, so

$$|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{MIN}(L_A)| \quad > \quad \epsilon_A. \tag{7}$$

Since all the entries in $\mathrm{Tr}_A(S)$ are non-negative, it follows from the definition that

$$\begin{aligned} & \mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|) \\ = \; & \frac{\sum_{\{B:\varphi(B)=\mathrm{F}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} \mathrm{Tr}_A(S)_i |S|}}{\sum_{\{B:\varphi(B)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} \mathrm{Tr}_A(S)_i |S|}} \\ \geq \; & \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} \qquad \text{[see below]}. \end{aligned}$$

The inequality holds because, as we observed before, the term in the numerator with the greatest exponent has exponent $\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|$. Its coefficient is at least $\min_B \Pr_{D,\vec{\alpha}}(B)$. The remaining terms in the numerator (if any) are nonnegative. Thus, the numerator is at least as large as $\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}$. There are 2^n terms in the denominator, each of which is at most $o^{|S|}$, since, as we observed earlier $\sum_i \mathrm{Tr}_A(S)_i = 1$ (since S is compatible with A). Thus, the denominator is at most $2^n o^{|S|}$.

Choose an arbitrary $\vec{d} \in \mathrm{OPT}(L_A(\varphi))$, and let $S_{|S|,\vec{d},A}$ be the approximation of Lemma B.20. For brevity, set $\vec{d'} = \mathrm{Tr}_A(S_{|S|,\vec{d},A})$. Then by Lemma B.20, $|\mathrm{maxp}_{\varphi,A}(\vec{d'}) - \mathrm{maxp}_{\varphi,A}(\vec{d})| < 2n/|S|$. So if $|S| > k_{\epsilon,A} \geq 4n/\epsilon_A$, then $|\mathrm{maxp}_{\varphi,A}(\vec{d'}) - \mathrm{maxp}_{\varphi,A}(\vec{d})| < \epsilon_A/2$. Since $\vec{d} \in \mathrm{OPT}(L_A)$, we have that $\mathrm{maxp}_{\varphi,A}(\vec{d}) = \mathrm{MIN}(L_A)$, so $|\mathrm{maxp}_{\varphi,A}(\vec{d'}) - \mathrm{MIN}(L_A)| \leq \epsilon_A/2$. Together with (7), Now using (7) and applying the triangle inequality gives us that

$$\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d'}) > \epsilon_A/2. \tag{8}$$

Much as above, we can show that

$$\begin{aligned} & \mathrm{cf}_A(\varphi, \vec{d'}, |S|) \\ = \; & \frac{\sum_{\{B:\varphi(B)=\mathrm{F}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} d'_i |S|}}{\sum_{\{B:\varphi(B)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} d'_i |S|}} \\ \leq \; & \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d'})|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}}, \end{aligned}$$

where now the inequality follows because we have replaced every term $\Pr_{D,\vec{\alpha}}(B)$ in the numerator by 1 and there are at most 2^n of them, and the fact that $\Pr_{D,\vec{\alpha}}(A) o^{|S|}$ is one of the terms in the denominator and the rest are non-negative.

Note that

$$\begin{aligned} & \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d'})|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}} \leq \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} \\ \text{iff} \quad & \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} - \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d'})|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}} \geq 0 \\ \text{iff} \quad & \frac{\Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|} - 2^{2n} o^{\mathrm{maxp}_{\varphi,A}(\vec{d'})|S|}}{\Pr_{D,\vec{\alpha}}(A) 2^n o^{|S|}} \geq 0 \\ \text{iff} \quad & \Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S| - \mathrm{maxp}_{\varphi,A}(\vec{d'})|S|} - 2^{2n} \geq 0 \\ \text{iff} \quad & (\mathrm{maxp}_{\varphi,A}(\underset{A}{\mathrm{Tr}}(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d'}))|S| \geq \log_o \frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B)}. \end{aligned}$$

By assumption, $|S| > \frac{2}{\epsilon_A}\log_o \frac{2^n 2^n}{\Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B)}$; by (8), $(\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d'})) > \epsilon_A/2$. Thus, $\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d'}))|S| \geq \log_o \frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(A)\min_B \Pr_{D,\vec{\alpha}}(B)}$, so By Lemma B.10,

$$\mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|) = \mathrm{cf}(\varphi, S) \text{ and } \mathrm{cf}_A(\varphi, \vec{d'}, |S|) = \mathrm{cf}(\varphi, S_{|S|,\vec{d},A}).$$

Next note that if

$$\frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d})|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}} \leq \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}}.$$

$$\mathrm{cf}(\varphi, S_{|S|,\vec{d},A}) < \mathrm{cf}(\varphi, S),$$

and so by Lemma B.4,

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S_{|S|,\vec{d},A}) > \Pr_{D,\vec{\alpha}}(\varphi \mid S),$$

so S is not optimal, as desired. □

Moreover, unless the sequence in question is short, any optimal sequence of test outcomes must be compatible with an LP that actually attains the minimax power.

Lemma B.22. *There exists a constant k_0, depending only on φ, D and $\vec{\alpha}$, such that if a sequence S of length $|S| \geq k_0$ is compatible with an assignment A, then either S is not optimal or A is relevant.*

Proof. Follows from a small modification to the proof of Lemma B.21. Rather than assuming (6), we simply assume that A is not relevant. Let B be an arbitrary relevant assignment; then this implies

$$\mathrm{MIN}(L_A) - \mathrm{MIN}(L_B) = \epsilon_2 > 0$$

and hence, as $\mathrm{Tr}_A(S)$ is a feasible point of $L_A(\varphi)$,

$$\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{MIN}(L_B) > \epsilon_2/2 =: \epsilon_1$$

can be used in (8)'s stead.

Then, instead of comparing S to a sequence $S_{|S|}(\vec{d}, A)$ synthesised from a solution point $\vec{d} \in \mathrm{OPT}(L_A)$, we take a sequence $S_{|S|}(\vec{d}, B)$ for a solution point $\vec{d} \in \mathrm{OPT}(L_B)$ of a relevant LP L_B. As before, we conclude that if $|S| > \max\{4n/\epsilon_1, k_2\}$ (k_2 defined as in the preceding proof, but in terms of the alternative choice of ϵ_1 here), a set of inequalities holds that together implies S is not optimal, and k_ϵ only depends on D, $\vec{\alpha}$ and the structure of φ (partially via our new ϵ_1). □

With these pieces, we can finally prove Theorem B.16.

Proof (of Theorem B.16). Suppose, by way of contradiction, that the antecedent of Theorem B.16 holds, but φ does not exhibit RI. By Proposition 4.7, there exists an open-minded product distribution D and accuracy vector α such that for all negligible functions f and constants $c > 0$, there exists an (f, c)-bad test-outcome sequence optimal for φ, D, and $\vec{\alpha}$. So by Lemma B.18, for all negligible functions f and constants $c > 0$, there exists an infinite sequence $\{S_k\}$ of (f, c)-bad test-outcome sequences that are optimal for φ, D, and $\vec{\alpha}$ and are of increasing length. Thus,

for all k, there are no variables $v_j \geq_\varphi v_i$ such that v_j is tested at most $f(|S_k|)$ times, but v_i is tested at least $c|S_k|$ times. (9)

We can assume without loss of generality that all the sequences S_k are compatible with the same assignment A, since there must be an assignment A that infinitely many of the sequences S_k are compatible with, and we can consider the subsequence consisting just of these test-outcomes sequences that are compatible with A. Moreover, by Lemma B.22, we can assume that A is relevant, since all but finitely many of the S_k must be sufficiently long.

Let δ be the function of Lemma B.21 and let C be the constant that is assumed to exist in the statement of Theorem B.16. Define f by taking $f(k) = \delta(k)k$. Since $\lim_{k\to\infty} f(k)/k = \lim_{k\to\infty} \delta(k) = 0$, f is negligible. By the above, there exists an sequence $\{S_k\}$ of $(f, C/2)$-bad test-outcome sequences optimal for φ, D, and $\vec{\alpha}$, all compatible with a relevant truth assignment A.

Let k_1 be sufficiently large that $\delta(k) < C/2$ for all $k > k_1$. By Lemma B.21, for all $k > k_1$, we must have

$$\|\vec{d} - \mathrm{Tr}_A(S_k)\|_1 < \delta(k) < C/2$$

for some solution $\vec{d}$ to the LP $L_A(\varphi)$. Since A is relevant by construction, the assumptions of the theorem guarantee that there exist i and j such that $v_i \leq_\varphi v_j$, $d_i > C$, and $d_j = 0$. Since $\|\vec{d} - \mathrm{Tr}_A(S_k)\|_1 < \delta(|S_k|)$, it follows that $(\mathrm{Tr}_A(S_k))_i > C - \delta(|S_k|) > C/2$ and $(\mathrm{Tr}_A(S_k))_j < \delta(|S_k|)$. Since each sequence S_k is compatible with A, for each variable v_h, $n^+_{S_k,A,h}$ is just the number the number of times that v_h is tested in S_k, so $(\mathrm{Tr}_A(S_k))_h$ is the number of times that v_h is tested divided by $|S_k|$. This means that we have a contradiction to (9). □

B.3 LP lower bound for rational inattention

Theorem B.16 gives us a criterion that is sufficient to conclude that a given formula exhibits rational inattention: there exists a C such that for all relevant assignments A, all points $\vec{c} = (c_1, \ldots, c_n, m) \in \mathrm{OPT}(L_A(\varphi))$ have entries c_i and c_j such that $v_i \leq_\varphi v_j$, $c_i > C$ and $c_j = 0$. We call this property P_C, and write $P_C(\vec{c})$ if $\vec{c}\, P_C$ satisfies it.

To compute how many formulae exhibit RI, we want an efficient algorithm that evaluates P_C. Unfortunately, the problem solved by standard linear programming algorithms is that of finding *some* point inside the solution polytope. A general way of checking if all points in $\mathrm{OPT}(L)$ for an LP L satisfy a property P is to separately determine the minimum m^+ of the objective function among all feasible points of L that satisfy P and the minimum m^- among all feasible points that don't. Then if $m^+ < m^-$, it follows that all points in $\mathrm{OPT}(L)$ satisfy P; similarly, if $m^- < m^+$, it follows that no points in $\mathrm{OPT}(L)$ satisfy P. Finally, if $m^+ = m^-$, then some points in OPT satisfy P and other points do not.

In general, the subset of feasible points that satisfy P may not be a convex polytope, so it may not be possible to use linear programming to determine m^+ and m^-. Indeed, the property that we care about, that is, the existence of indices i and j such that $v_i \leq_\varphi v_j$, $c_i > C$, and $c_j = 0$, is not even closed under convex combinations, let alone expressible as a set of linear inequalities. For example, if $v_i \leq_\varphi v_j$ and $v_j \leq_\varphi v_i$ are two variables of equal relevance, then the points $(\ldots, 0, \ldots, 0.2, \ldots)$ and $(\ldots, 0.2, \ldots, 0, \ldots)$ (the filled-in entries correspond to coordinates i and j) satisfy the property for i and j, but their average $(\ldots, 0.1, \ldots, 0.1, \ldots)$ does not. However, for fixed i and j, the condition that $c_i > C$ and $c_j = 0$ can be imposed easily by adding the two inequalities in question to the LP. The set of points that satisfy the existentially quantified condition therefore can be covered by a $O(n^2)$-sized family of convex polytopes, over which we can minimise m as a linear program, and determine the overall minimum m^+ by taking the minimum over the individual minima.

Definition B.23. For all variables $v_i \neq v_j$ with $v_i \leq_\varphi v_j$, define

$$L^+_{A,i,j}(\varphi, C) = L_A(\varphi) \cup \{c_j = 0, c_i \geq C\}$$

(so, roughly speaking, in solutions to $L^+_{A,i,j}(\varphi, C)$, variable v_j is ignored while v_i is tested in a constant fraction of the tests). □

Clearly, $\bigcup L^+_{A,i,j}(\varphi, C) = \{\vec{p} \in \mathit{Feas}(L_A) \mid P_C(\vec{p})\}$, so $\min_{i,j} \mathrm{MIN}(L^+_{A,i,j}) = m^+$.

To determine m^-, we need to similarly cover the set of points on which P_C is *not* satisfied with convex polytopes. The negation of P_C is

$$\forall j(\, c_j = 0 \Rightarrow \forall i(\, v_i \leq_\varphi v_j \Rightarrow c_i < C)).$$

Definition B.24. The *attentive set* (based on L_A, i and C) $S^-_{A,i}(\varphi, C)$ is the subset of the feasible set of L_A where those variables v_j that are at most as important as v_i (that is, $v_j \leq_\varphi v_i$) are "mostly ignored", and the variables that are more important are not ignored:

$$S^-_{A,i}(\varphi, C) = \{(c_1, \ldots, c_n, m) \in \mathit{Feas}(L_A(\varphi)) \mid c_j < C \text{ for all } j \in I_i,\, c_j > 0 \text{ for all } j \notin I_i\}.$$

□

Intuitively, v_i is the "last" variable (in the $\leq_\varphi$ ordering) such that $c_i = 0$. Thus, for all j with $v_i \leq_\varphi v_j$, we must have $c_j < C$, and for all j with $v_j <_\varphi v_i$, we must have $c_i > 0$. It is easy to see that $\bigcup S^-_{A,i}(\varphi, C) \supseteq \{\vec{p} \in \mathit{Feas}(L_A) \mid \neg P_C(\vec{p})\}$. Unfortunately, the definition of $S^-_{A,i}(\varphi, C)$ involves some strict inequalities, so we cannot use LP techniques to solve for m^- in the same way as we solved for m^+.

We are ultimately interested in whether $m^+ < m^-$. This is the case if there is no point in $S^-_{A,i}$ such that $m \leq m^+$; that is, if

$$T^-_{A,i}(\varphi, C, m^+) = \{(c_1, \ldots, c_n, m) \in S^-_{A,i}(\varphi, C) \mid m \leq m^+\} = \varnothing$$

for all A and i. This is a set that is defined by linear inequalities, and so we can use standard results to determine if it contains any point, e.g. [cite?] or by adding a slack variable to all strict inequalities, maximising it with LP techniques and checking that the maximum is nonzero.

Theorem B.25. *Choose an arbitrary C. If m^+ is the minimum of $\mathrm{MIN}(L^+_{A,i,j}(\varphi, C))$ over all inattentive LPs $L^+_{A,i,j}(\varphi, C)$ and all sets $T^-_{A,i}(\varphi, C, m^+)$ are empty, then φ exhibits rational inattention.*

Proof. As explained above, the sets $T^-_{A,i}$ being empty implies that there is no point satisfying $\neg P_C$ and attaining a max-power of $m \leq m^+$. At the same time, m^+ being the minimum over all inattentive LPs means that the minimum of m over points satisfying P_C in any L_A is m^+. Therefore, m^+ is the minimax power, and all solution points of relevant LPs satisfy P_C with our arbitrary choice of C. Hence φ exhibits RI per Theorem B.16. □

Corollary B.26. *We can compute a sufficient condition for the n-variable formula φ to exhibit RI by solving $2^n O(n^2)$ LPs with $O(2^n)$ inequalities each, namely the $O(n^2)$ inattentive LPs and the $O(n)$ attentive LPs associated with each of the 2^n assignments.*

C Proof of Theorem 5.2

To prove Theorem 5.2, we first need a definition and some lemmas.

Definition C.1. Given a probability on some space S and an event $E \subseteq S$, define the *bias* of E to be

$$Q(E) = \Pr_{D,\vec{\alpha}}(E) - \frac{1}{2}.$$

□

The following lemma will be helpful in establishing the complexity of formulae.

Lemma C.2. *If, for all test-outcome sequences S, there exists a test-outcome sequence S' such that $|S'| = |S|$ and*

$$|\,\mathrm{Q}(\varphi|S')| \geq |\,\mathrm{Q}(\psi|S)|,$$

then

$$\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi) \leq \mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\psi).$$

Proof. Suppose that $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\psi) = k$. Then there must be some strategy σ for $G(\psi, D, k, \vec{\alpha}, g, b)$ that has positive expected payoff. Analogously to the argument in the proof of Proposition 3.1, there must therefore be some test-outcome sequence S of length k that is observed by σ with positive probability such that the expected payoff of making the appropriate guess is positive. By Lemma 3.3, $|\,\mathrm{Q}(\psi|S)| > q$.

Since $|\,\mathrm{Q}(\varphi|S')| \geq |\,\mathrm{Q}(\psi|S)|$ by assumption, there must exist a test-outcome sequence S' such $|\,\mathrm{Q}(\varphi|S')| > q$. Let σ' be the strategy for the game $G(\varphi, D, k, \vec{\alpha}, g, b)$ that tests the same variables that are tested in S', and makes the appropriate guess if and only if S' is in fact observed. By Lemma 3.3, a guess with positive expected payoff can be made if S' is observed, which it is with positive probability. So σ' has positive expected payoff, and hence $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi)$ is at most k. □

The following consequence of the definition of a product distribution will be very useful in this chapter.

Lemma C.3. *For all product distributions D, accuracy vectors $\vec{\alpha}$, assignments A and test-outcome sequences S,*

$$\Pr_{D,\vec{\alpha}}(A \mid S) = \prod_{i=1}^{n} \Pr_{D,\vec{\alpha}}(v_i = A(v_i) \mid S).$$

Proof. The event A is the intersection of the n events $v_i = A(v_i)$, and since D is a product distribution, the events are independent of each other. □

Definition C.4. Let v be a variable and φ be a Boolean formula in variables including v. The *projections of φ along v are* the two formulae $\varphi|_{v=\mathrm{T}}$ and $\varphi|_{v=\mathrm{F}}$, where $\varphi|_{v=x}(A)$ denotes the formula that results from replacing all occurrences of v in φ by x.

Note that $\varphi \equiv (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \varphi|_{v=\mathrm{F}})$. We call φ *antisymmetric in v* if $\varphi|_{v=\mathrm{T}} \equiv \neg\varphi|_{v=\mathrm{F}}$. □

Lemma C.5. *Suppose D is a product distribution. Then, given a sequence of test outcomes S, the projection of a formula $\varphi|_{v_i=b}$ has the same conditional probability on S as φ additionally conditioned on $v_i = b$, that is:*

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S, v_i = b) = \Pr_{D,\vec{\alpha}}(\varphi|_{v_i=b} \mid S).$$

Proof. We have

$$\begin{aligned}
\Pr_{D,\vec{\alpha}}(\varphi|_{v_i=b} \mid S) &= \sum_{\{A:\varphi|_{v_i=b}(A)=T\}} \Pr_{D,\vec{\alpha}}(A \mid S) \\
&\overset{C.3}{=} \sum_{\{A:\varphi|_{v_i=b}(A)=T\}} \prod_{1\leq j\leq n} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S).
\end{aligned}$$

Expanding the definition of the projection, this is

$$\begin{aligned}
&= \sum_{\{A:\varphi(A)=T,A(v_i)=b\}} \prod_{1\leq j\leq n} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S) + \sum_{\{A:\varphi(A[v_i\mapsto b])=T,A(v_i)=\neg b\}} \prod_{1\leq j\leq n} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S) \\
&= \sum_{\{A:\varphi(A)=T,A(v_i)=b\}} \left(\prod_{1\leq j\leq n} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S) + \prod_{1\leq j\leq n} \Pr_{D,\vec{\alpha}}(v_j = A[v_i \mapsto \neg b](v_j) \mid S) \right) \\
&= \sum_{\{A:\varphi(A)=T,A(v_i)=b\}} (\Pr_{D,\vec{\alpha}}(v_i = b \mid S) + \Pr_{D,\vec{\alpha}}(v_i = \neg b \mid S)) \prod_{1\leq j\leq n, j\neq i} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S) \\
&= \sum_{\{A:\varphi(A)=T,A(v_i)=b\}} \prod_{1\leq j\leq n, j\neq i} \Pr_{D,\vec{\alpha}}(v_j = A(v_j) \mid S) \\
&= \Pr_{D,\vec{\alpha}}(\varphi, v_i = b \mid S) / \Pr_{D,\vec{\alpha}}(v_i = b \mid S) \\
&= \Pr_{D,\vec{\alpha}}(\varphi \mid S, v_i = b) \text{ (Bayes).}
\end{aligned}$$

□

The notion of antisymmetry has the useful property that it is possible to "antisymmetrise" a formula along a particular variable without interfering with its antisymmetry along others:

Lemma C.6. *If φ is antisymmetric in a variable $y \neq v$ and $\varphi' = (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \neg\varphi|_{v=\mathrm{T}})$, then φ' is also antisymmetric in y.*

Proof. Using the fact that φ is antisymmetric in y, for all truth assignments A, we have

- $\varphi(A[y \mapsto \mathrm{T}]) = \neg\varphi(A[y \mapsto \mathrm{F}])$ and
- $\varphi'(A) = \begin{cases} \varphi(A[v \mapsto \mathrm{T}]) & \text{if } A(v) = \mathrm{T} \\ \neg\varphi(A[v \mapsto \mathrm{T}]) & \text{if } A(v) = \mathrm{F}; \end{cases}$

Thus,

$$\begin{aligned} \varphi'(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{T}]) &= \varphi(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{T}]) \\ &= \neg\varphi(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{F}]) \\ &= \neg\varphi'(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{F}]) \end{aligned}$$

and

$$\begin{aligned} \varphi'(A[v \mapsto \mathrm{F}, y \mapsto \mathrm{T}]) &= \neg\varphi(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{T}]) \\ &= \neg\neg\varphi(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{F}]) \\ &= \varphi(A[v \mapsto \mathrm{T}, y \mapsto \mathrm{F}]) \\ &= \neg\varphi'(A[v \mapsto \mathrm{F}, y \mapsto \mathrm{F}]), \end{aligned}$$

as required. □

Define $V(\varphi)$, the number of variables a formula φ is *not* antisymmetric in, as

$$V(\varphi) = |\{v : \varphi \not\equiv (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \neg\varphi|_{v=\mathrm{T}})\}|.$$

Proposition C.7. *The only formulae φ in the n variables $v_1, \ldots, v_n$ for which $V(\varphi) = 0$ are equivalent to either $\bigoplus_{i_1}^{n} v_i$ or $\neg\bigoplus_{i=1}^{n} v$.*

Proof. By induction on n. If $n = 1$, then it is easy to check that both v_1 and $\neg v_1$ are antisymmetric. Suppose that $n > 1$ and φ is antisymmetric in $v_1, \ldots, v_n$. Since $\varphi \equiv (v_n \wedge \varphi|_{v_n=\mathrm{T}}) \vee (\neg v_n \wedge \varphi|_{v_n=\mathrm{F}})$ and φ is antisymmetric in v_n, we have that

$$\varphi \equiv (v_n \wedge \varphi|_{v_n=\mathrm{T}}) \vee (\neg v_n \wedge \neg\varphi|_{v_n=\mathrm{T}}) = v_n \oplus \varphi|_{v_n=\mathrm{T}}.$$

It is easy to see that $\varphi|_{v_n=\mathrm{T}}$ mentions only the variables $v_1, \ldots, v_{n-1}$ and is antisymmetric in each of them, so by the induction hypothesis φ is equivalent to either $\oplus_{i=1}^{n-1} v_i$ or $\neg(\oplus_{i=1}^{n-1} v_i)$. □

We can now prove Theorem 5.2.

Proof of Theorem 5.2. We show by induction on $V(\varphi)$ that for all formulae φ, there exists a formula φ_0 with $V(\varphi_0) = 0$ such that $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}\,\varphi \leq \mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}\,\varphi_0$. By Proposition C.7, φ_0 must be equivalent to either $\oplus_{i=1}^{n-1} v_i$ or $\neg(\oplus_{i=1}^{n-1} v_i)$.

If $V(\varphi) = 0$, then we can just take $\varphi_0 = \varphi$. Now suppose that $V(\varphi) > 0$. Then must exist some variable v such that $\varphi|_{v=\mathrm{T}} \neq \varphi|_{v=\mathrm{F}}$. Let

$$\varphi' = (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \neg\varphi|_{v=\mathrm{T}}).$$

Note for future reference that

$$\varphi'|_{v=\mathrm{T}} = \varphi'|_{v=\mathrm{T}} \text{ and } \varphi'|_{v=\mathrm{F}} = \neg\varphi|_{v=\mathrm{T}}. \tag{10}$$

By Lemma C.6, if φ is antisymmetric in a variable $v' \neq v$, then so is φ'. In addition, φ' is antisymmetric in v. Thus, $V(\varphi') < V(\varphi)$. Thus, if $\mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi) \leq \mathrm{cpl}_{\mathrm{D,q},\vec{\alpha}}(\varphi')$, then the result will follow from the induction hypothesis. To do this, by Lemma C.2, it suffices to show that for all test-outcome sequences S_1, there exists a sequence S of the same length as S_1 such that

$$|\,\mathrm{Q}(\varphi|S)| \geq |\,\mathrm{Q}(\varphi'|S_1)|.$$

Given an arbitrary test-outcome sequence S_1, let $p = \Pr_{D,\vec{\alpha}}(v = \mathrm{T} \mid S_1)$. Because the events $\{v = \mathrm{T}, v = \mathrm{F}\}$ partition the sample space, we have

$$\Pr_{D,\vec{\alpha}}(\varphi' \mid S_1) = p \Pr_{D,\vec{\alpha}}(\varphi' \mid S_1, v = \mathrm{T}) + (1-p) \Pr_{D,\vec{\alpha}}(\varphi' \mid S_1, v = \mathrm{F}),$$

and so

$$\begin{aligned}\Pr_{D,\vec{\alpha}}(\varphi' \mid S_1) &= p\Pr_{D,\vec{\alpha}}(\varphi' \mid S_1, v=\mathrm{T}) + (1-p)\Pr_{D,\vec{\alpha}}(\varphi' \mid S_1, v=\mathrm{F}) \\ &\overset{C.5}{=} p\Pr_{D,\vec{\alpha}}(\varphi'|_{v=\mathrm{T}} \mid S_1) + (1-p)\Pr_{D,\vec{\alpha}}(\varphi'|_{v=\mathrm{F}} \mid S_1) \\ &= p\Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p)\Pr_{D,\vec{\alpha}}(\neg\varphi|_{v=\mathrm{T}} \mid S_1) \text{ [by (10)].}\end{aligned}$$

Note that if $\Pr_{D,\vec{\alpha}}(\varphi_1|S_1) = p\Pr_{D,\vec{\alpha}}(\varphi_1|S_2) + (1-p)\Pr_{D,\vec{\alpha}}(\varphi_3|S_3)$, then it is easy to check that

$$\mathrm{Q}(\varphi_1|S_1) = \Pr_{D,\vec{\alpha}}(\varphi_1|S_1) - 1/2 = p\,\mathrm{Q}(\varphi_1|S_2) + (1-p)\,\mathrm{Q}(\varphi_3|S_3).$$

Moreover,

$$\mathrm{Q}(\neg\varphi|S) = \Pr_{D,\vec{\alpha}}(\neg\varphi|S) - 1/2 = 1/2 - \Pr_{D,\vec{\alpha}}(\varphi|S) = \mathrm{Q}(\varphi|S).$$

Thus,

$$\mathrm{Q}(\varphi'|S_1) = p\,\mathrm{Q}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p)\,\mathrm{Q}(\neg\varphi|_{v=\mathrm{T}} \mid S_1) = p\,\mathrm{Q}(\varphi|_{v=\mathrm{T}} \mid S_1) - (1-p)\,\mathrm{Q}(\varphi|_{v=\mathrm{T}} \mid S_1). \tag{11}$$

Set $S_2 = S_1[v \approx \mathrm{F} \leftrightarrow v \approx \mathrm{T}]$, that is, the sequence that is the same as S_1 except that all test outcomes of v are flipped in value. We now show that either $|\,\mathrm{Q}(\varphi|S_1)| \geq |\,\mathrm{Q}(\varphi'|S_1)|$ or $|\,\mathrm{Q}(\varphi|S_2)| \geq |\,\mathrm{Q}(\varphi'|S_1)|$.

Since $\varphi|_{v=\mathrm{T}}$ does not mention v, $\mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_1) = \mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_2)$ and likewise for $\varphi|_{v=\mathrm{F}}$. Since $\varphi \equiv (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \varphi|_{v=\mathrm{F}})$, we have (using an argument similar to that above)

$$\mathrm{Q}(\varphi|S_1) = p\,\mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_1) + (1-p)\,\mathrm{Q}(\varphi|_{v=\mathrm{F}}|S_1) \tag{12}$$

and

$$\mathrm{Q}(\varphi|S_2) \quad = \Pr_{D,\vec{\alpha}}(v=\mathrm{T}|S_2)\,\mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_2) + \Pr_{D,\vec{\alpha}}(v=\mathrm{F}|S_2)\,\mathrm{Q}(\varphi|_{v=\mathrm{F}}|S_2).$$

Suppose v is the ith variable v_i. Let A be an arbitrary assignment such that $A(v_i) = T$ and A' such that $A(v_i) = F$. Then $n^+_{S,A,i}$ is independent of the choice of A for all S, and likewise $n^+_{S,A',i}$. Let $r_1 = o_i^{n^+_{S_1,A,i}}$ be the factor associated with v_i in all $r_{D,\vec{\alpha}}(A, S_1)$ where $A(v_i) = T$, and $r_2 = o_i^{n^+_{S_1,A',i}}$ be the factor associated with v_i in all $r_{D,\vec{\alpha}}(A, S_1)$ where $A(v_i) = F$. Also, let

$$R = \sum_{A:A(v_i)=T} \prod_{j=1,j\neq i}^{n} o_i^{n^+_{S_1,A,i}}.$$

Then by Lemma B.1 and uniformity of the prior (so all terms of the form $\Pr_{D,\vec{\alpha}}(A)$ cancel),

$$\Pr_{D,\vec{\alpha}}(v_i = T \mid S_1) = \frac{r_1 R}{r_1 R + r_2 R}.$$

For S_2, the remainder R is the same, while $n^+_{S_1,A,i} = n^+_{S_2,A',i}$ and vice versa, and so likewise

$$\Pr_{D,\vec{\alpha}}(v_i = T \mid S_2) = \frac{r_2 R}{r_1 R + r_2 R}.$$

Therefore,

$$\Pr_{D,\vec{\alpha}}(v_i = T \mid S_1) = 1 - \Pr_{D,\vec{\alpha}}(v_i = T \mid S_2)$$

and hence

$$= (1-p)\,\mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_1) + p\,\mathrm{Q}(\varphi|_{v=\mathrm{F}}|S_1), \tag{13}$$

To simplify notation, let $x = \mathrm{Q}(\varphi|_{v=\mathrm{T}}|S_1)$ and let $y = \mathrm{Q}(\varphi|_{v=\mathrm{F}})$. By (11), (12), and (13), we want to show that either $|px + (1-p)y| \geq |px - (1-p)x|$ or $|(1-p)x + py| \geq |px - (1-p)x|$. So suppose that $|px + (1-p)y| < |px - (1-p)x|$. We need to consider four cases: (1) $p \geq 1/2$, $x \geq 0$; (2) $p \geq 1/2$, $x < 0$; (3) $p < 1/2$, $x \geq 0$; and (4) $p < 1/2$, $x < 0$. For (1) , note that if $p \geq 1/2$ and $x \geq 0$, then $0 \leq px - (1-p)x \leq px$. We must have $y < -x$, for otherwise $px + (1-p)y \geq px - (1-p)x$. But then $py + (1-p)x < -(px - (1-p)x)$, so $|py + (1-p)x| > |px - (1-p)x|$.

For (2), note that if $p \geq 1/2$ and $x < 0$, then $px - (1-p)x < 0$. We must have $y > -x$, for otherwise $px + (1-p)y \leq px - (1-p)x$, and $|px + (1-p)y| \geq |px - (1-p)x|$. But then $py + (1-p)x > -px + (1-p)x$, so $|py + (1-p)x| > |px - (1-p)x|$.

The arguments in cases (3) and (4) are the same as for (1) and (2), since we can simply replace p by $1-q$. This gives us identical inequalities (using q instead of p), but now $q > 1/2$. □