Information Acquisition Under Resource Limitations in a Noisy Environment*

MATVEY SOLOVIEV, Computer Science Department, Cornell University, USA
JOSEPH Y. HALPERN, Computer Science Department, Cornell University, USA

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*, that is, when it is rational to ignore pertinent information (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.

CCS Concepts: • **Computing methodologies** → **Sequential decision making**; *Planning under uncertainty*; • **Theory of computation** → *Algorithmic game theory*.

ACM Reference Format:
Matvey Soloviev and Joseph Y. Halpern. 2022. Information Acquisition Under Resource Limitations in a Noisy Environment. 1, 1 (July 2022), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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. Our framework is motivated by a variety of decision problems in which multiple noisy signals are available for sampling, such as the following:

- An animal must decide whether some food is safe to eat. We assume that "safe" is characterised by a Boolean formula φ, which involves variables that describe (among other things) the 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

*A preliminary version of this paper appeared in *Proc. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 6443–6450.

Authors' addresses: Matvey Soloviev, Computer Science Department, Cornell University, 107 Hoy Rd, Ithaca, NY, 14853, USA, msoloviev@cs.cornell.edu; Joseph Y. Halpern, Computer Science Department, Cornell University, 107 Hoy Rd, Ithaca, NY, 14853, USA, halpern@cs.cornell.edu.

XXXX-XXXX/2022/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

payoff of 0 if the agent guesses "false" or does not guess, since both choices amount to not eating the food.

- A quality assurance team needs to certify a modular product, say a USB memory stick, or send it back to the factory. Some subsystems, such as the EEPROM cells, are redundant to an extent, and a limited number of them not working is expected and does not stop the product from functioning. Others, such as the USB controller chip, are unique; the device will not work if they are broken. Whether the device is good can be expressed as a Boolean combination of variables that describe the goodness of its components. Time and financial considerations allow only a limited number of tests to be performed, and tests themselves have a probability of false negatives and positives. What parts should be tested and how often?
- A data scientist wants to perform a complex query on a very big database. A certain error rate is acceptable; in any case, executing the query exactly is infeasible with the available hardware. The selection criterion itself is a Boolean combination of some atomic predicates on the entries of the database, which can be evaluated only using heuristics (which are essentially probabilistic algorithms). Given a query that asks for rows that, for instance, satisfy the criterion $P_1 \wedge (P_2 \vee P_3)$ in three predicates P_i, which heuristics should be run and how often should they be run to attain the desired error rate?

We are interested in optimal strategies for each of these problems; 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 [10, 14]. 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. (Roughly 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 certain 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. Our notion is based on the minimum number of tests required to have a chance of making a reasonable guess regarding whether the formula is true. 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., [3, 7, 9]).

Organisation. The rest of the paper is organized as follows. In Section 2, we formally define the games that we use to model our decision problem and analyse the optimal strategies for a simple example. The detailed calculations for this example can be found in Appendix A. In Section 3, we

look at the problem of determining optimal strategies more generally. We discuss the difficulty of this problem and analyse a simple heuristic, developing our understanding of the connection between payoffs and certainty in the process. In Section 4, we formally define rational inattention and discuss the intuition behind our definition. After considering some examples of when RI occurs under our definition, we show that there is a close connection between rational inattention and particular sequences of observations (*optimal test outcome sequences*) that may occur while testing. We use this connection to obtain a quantitative estimate of how common RI is in formulae involving up to 10 variables. The theory behind this estimate is presented in Appendix B, where we relate the optimal test outcome sequences to the solution polytope of a particular linear program (LP). While we are not aware of any explicit connections, our method should be seen in a broader tradition of applying LPs to decision problems such as multi-armed bandits [1], and may be of independent interest for the analysis of information acquisition. Finally, in Section 5, we introduce our notion of test complexity, prove that XORs are the formulas of greatest test complexity (the details of the proof are in Appendix C), and discuss the connections to various other notions of formula complexity in the cognitive and computational science literature.

2 INFORMATION-ACQUISITION GAMES

We model the *information-acquisition game* as a single-player game against nature, that is, one in which actions that are not taken by the player are chosen at random. The game is characterised by five parameters:

- a Boolean formula φ over 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 (possibly 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 A as a function from variables to truth values ($\{T, F\}$); we can also view a formula 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] Thus, outcomes are independent, conditional on a truth assignment. 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 his guess coincides with the actual truth value of φ on assignment A (i.e., his guess is correct) and b (bad) if his guess is wrong. It is occasionally useful to think of a formula φ as a function from assignments to truth values; we thus occasionally write $\varphi(A)$ to denote the truth value of φ under truth assignment A. A strategy for an agent in this game can be seen as a pair of functions: one that determines which test the agent performs after observing a given sequence of test outcomes of length $< k$, and one that determines whether to make a guess and, if so, which guess to make, given all k test outcomes.

[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.

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 Appendix A, the strategy "test v_1 and then v_2, then guess T if both tests came out T" is strictly worse than these two; in fact, its expected payoff is negative! This was (to us, at least) surprising: among other things, it implies that optimal strategies do not form a convex set, so a convex combination of two optimal testing strategies is not necessarily optimal.

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 $v_1 \vee v_2$.

For our results, we need to analyze the probability of various events related to the game. Many of the probabilities that we care about depend on only a few parameters of the game. Formally, we put a probability on *histories* of an information-acquisition game. A history is a tuple of the form (A, S, a), where A is the assignment of truth values to the n variables chosen by nature, $S = (v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k)$ is a *test-outcome sequence* in which $v_{i_j} \approx b_j$ 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 for this game then induce a probability $\Pr_{G,\sigma}$ on this sample space.

Example 2.2. In Example 2.1, $\Pr_{G,\sigma}(\varphi)$ is $3/4$, as we know only that there is a probability of $3/4$ that nature picked a satisfying assignment. After observing a single test outcome suggesting that v_1 is false, the posterior probability $\Pr_{G,\sigma}(\varphi \mid (v_1 \approx F))$ drops to $5/8$. If the same test is performed and the outcome is again F, the posterior drops further to $\Pr_{G,\sigma}(\varphi \mid (v_1 \approx F, v_1 \approx F)) = 11/20$. □

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 $\Pr_{G,\sigma}(\varphi \mid S)$, so we write $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$. 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[2] 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

[2]For the one-player games that we are considering, a game tree is a graph whose nodes consist of all valid partial sequences of actions in the game, including the empty sequence, and two nodes have an edge between them if they differ by appending one action.

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, and hence grows exponentially in k even for a fixed number of variables n, quickly becoming infeasible.

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 there is a strategy that gets a positive expected payoff when the bound on tests is k 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, so it lies in a complexity class that is at least as hard as both NP and co-NP.

A simple heuristic (whose choice of variables is independent of φ) would be to simply test each variable in the formula 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 is small (note that k still can be large, since we can test a variable multiple times!), so this time complexity would be acceptable. However, this approach can be arbitrarily worse than the optimum. As we observed when discussing 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. The *threshold* associated with payoffs b, g is $q(b,g) = \frac{b+g}{2(b-g)}$. □

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 iff*

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

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 test-outcome sequences of length k partition the space of paths in the game tree, so we have

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

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, $\left|\Pr_{D,\vec{\alpha}}(\varphi \text{ is true} \mid S^*) - 1/2\right| > q(b,g)$.

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

4.1 Defining 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.[3] 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.

[3]We checked various other reasonable definitions experimentally; qualitatively, it seems that our results continue to hold for all the variants that we tested.

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 that we prove could be obtained for a probabilistic notion of relevance ordering.

Roughly speaking, φ exhibits RI if, for all optimal strategies σ for the game $G(\varphi, D, k, \vec{\alpha}, b, g)$, 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 the rational choice is always to guess $\varphi(A)$, regardless of the test outcomes.

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).[4] We now make these ideas precise.

Definition 4.2. A function $f : \mathbb{N} \to \mathbb{N}$ is *negligible* if $f(k) = o(k)$, that is, 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'

[4]One way to avoid these additional requirements is to modify the game so that performing a test 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.

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 $c > 0$). (Note that this additional requirement makes it harder for a variable to exhibit RI.)

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 $v_1, \ldots, v_n$ is a *product distribution* if $\Pr_D(A) = \Pr_D(v_1 = A(v_1)) \cdots \Pr_D(v_n = A(v_n))$ (where, for an arbitrary formula φ, $\Pr_D(\varphi) = \sum_{\{A:\, A(\varphi)=\mathrm{T}\}} \Pr_D(A)$). □

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. Similarly, 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 histories of the game, 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. □

Our definition of RI is quite strong; for instance, as we discuss at the end of Section 4, we could obtain a plausible weakening by requiring that a variable v' at least as relevant as a variable v is tested far less frequently than v in a set of histories with positive probability rather than in *all* histories. But even with our strong requirements, we find that rational inattention in the given strong sense is quite widespread.

To get an intuition for this definition of RI, we will first directly check whether some natural classes of formulae satisfy it.

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 false (for conjunctions) or that it is true (for disjunctions). 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 a XOR in more than one variable, an optimal strategy must test all of the variables about the same number of times, 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 those equivalent to one of the four conjunctions $\ell_1 \wedge \ell_2$ or 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.
5. For $n > 4$, formulae φ of the form $(v_1 \vee v_2) \wedge (v_3 \oplus \ldots \oplus v_n)$ exhibit RI. Optimal strategies split tests between v_1 and v_2, and try to establish that both are false, and hence that φ is. To establish that φ is true would require showing that the XOR is true. This, in turn, would require testing all of $v_3, \ldots, v_n$; since $n > 4$, there are at least three variables to test. As we noted earlier, an optimal strategy must test each of the variables in the XOR about the same number of times to establish the truth (or falsity) of the XOR. It thus requires signficantly more tests to gain a given level of confidence that the XOR is true (or false) than it does to gain that level of confidence that $v_1 \vee v_2$ is false. (In Section 5, we show that XORs are the formulae that we learn the least about in a given number of tests among all formulas with a fixed number of variables.) The variable v_1 determines whether is true in only φ in $1/4$ of the assignments (when the XOR is true, which it is in half the assignments, and v_2 is false); similarly for v_2. On the other hand, $v_3, \ldots, v_n$ determine the truth value of φ in $3/4$ of all assignments (all assignments where $v_1 \vee v_2$ is true). Thus, this family of formulae (and other similar families) satisfy an even stronger definition of RI, as a strictly less relevant variable is preferred. □

4.2 A sufficient criterion for rational inattention

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 $|S'| = |S|$. □

It turns out that for a formula to exhibit rational inattention, it is sufficient (but not necessary!) for just the optimal test-outcome sequences to be "inattentive", because we can set up the payoffs in such a way that only the very best test-outcome sequences ever become relevant (by possibly leading to a non-negative payoff). By doing this, we avoid having to deal with the complicated quantification over all histories in the definition of RI. With the appropriate payoffs, each history either has to end in no guess or contain an optimal test-outcome sequence. We will see that we can reason about optimal test-outcome sequences without having to worry about the structure of arbitrary optimal strategies.

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

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

Assume for now that $q^* > 0$. Since there are only finitely many test-outcome sequences of length k, there must be some ϵ with $q^* > \epsilon > 0$ sufficiently small such that for all S with $|S| = k$, $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| > q^* - \epsilon$ iff $|\Pr_{D,\vec{\alpha}}(\varphi|S) - \frac{1}{2}| = q^*$. Choose the payoffs b and g such that the threshold $q(b, g)$ is $q^* - \epsilon$. 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, the 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 establish 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 iff S was in fact the test-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.

It remains to address the case where $q^* = 0$, that is, the number of tests k is insufficient to learn anything about the truth value of the formula. In this case, we can simply set $b = -1$ and $g = 2$, and proceed as before, needing to show only that some strategy attains a positive payoff. Indeed, take σ_T to be a strategy that repeatedly tests some variable v and then guesses T regardless of outcomes. Since $q^* = 0$, so we have that $\Pr_{D,\vec{\alpha}}(\varphi \mid S) = 1/2$ for all test-outcome sequences, there is a probability $1/2$ of getting payoff -1 and a probability $1/2$ of getting payoff 2, so the expected payoff is positive. This completes the proof. □

Applying Proposition 4.7 to test whether a formula exhibits RI is not trivial. It is easy to show that all that affects $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ is the number of times that each variable is tested and the outcome of the test, not the order in which the tests were made. We do this formally in Appendix B.1; the following notation, which we use throughout the rest of this section and in Appendix B, implicitly assumes this fact.

Definition 4.8. (General notation)

- $o_i = \frac{1/2+\alpha_i}{1/2-\alpha_i}$. 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.
- $n^+_{S,A,i} = |\{j : S[j] = (v_i \approx A(v_i))\}|$. 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.
- $r_{D,\vec{\alpha}}(A, S) = \Pr_{D,\vec{\alpha}}(A) \prod_{\{i:v_i \text{ is in the domain of } A\}} o_i^{n^+_{S,A,i}}$ □

Using these definitions, we can define the following quantity, which forms the starting point of our approach. Its value represents an anticorrelate of conditional probability, while its definition packages up the notion that the ordering of a test-outcome sequence doesn't matter for easier use.

Definition 4.9. 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 4.10. $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > \Pr_{D,\vec{\alpha}}(\varphi \mid S')$ *iff* $\mathrm{cf}(\varphi, S) < \mathrm{cf}(\varphi, S')$.

Proof. See Appendix B.1. □

Example 4.11. 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 and the testing accuracy is the same for all variables, so $o_1 = \ldots = o_n = o$. This formula has four satisfying assignments, namely $\{TTT, TFF, TTF, FFF\}$ (letting xyz denote the assignment $\{v_1 \mapsto x, v_2 \mapsto y, v_3 \mapsto z\}$, for brevity). The other four assignments, namely $\{FFT, TFT, FTT, FTF\}$, make the formula false. For each assignment A, the corresponding summand $r_{D,\vec{\alpha}}(A, S)$ is $\Pr_{D,\vec{\alpha}}(A)$ times a factor of o for every test outcome in S that is compatible with A, where a test outcome $v_i \approx b$ is *compatible* with A if $b = A(v_i)$. For instance, the falsifying assignment FFT is compatible with $v_2 \approx F$ but not $v_1 \approx T$, so it gives rise to a summand of $\Pr_{D,\vec{\alpha}}(A) \cdot o$ in the numerator of the characteristic fraction of S. On the other hand, if A is the satisfying assignment TFF, then both $v_1 \approx T$ and $v_2 \approx F$ are compatible with A, yielding $\Pr_{D,\vec{\alpha}}(A) \cdot o^2$ in the denominator. Performing the same analysis for the other assignments and cancelling the common factors of $\Pr_{D,\vec{\alpha}}(A)$ (as the prior is uniform), we find that

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

For a more general example, suppose that $S = ((v_1 \approx T)^{c_1 k}, (v_2 \approx F)^{c_2 k}, (v_3 \approx F)^{c_3 k})$, where k is the total number of tests in S and real constants $0 \le c_1, c_2, c_3 \le 1$ with $c_1 + c_2 + c_3 = 1$ representing the fraction of times that each of the three test outcomes occurs in it. Then

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

□

In the second example above, the characteristic fraction of S depends only on the integers $c_1 k$, $c_2 k$, and $c_3 k$. The factor k is common to all the exponents, and so we can pull it out to rewrite $\mathrm{cf}(\varphi, S)$ as a fraction of sums of powers of o^k :

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

The key point is that we can view $\mathrm{cf}(\varphi, S)$ as a function of the vector (c_1, c_2, c_3) that describes the relative makeup of the test-outcome sequence and a parameter o^k that depends on the test accuracy and the number of tests k. This can be shown to be true in general. Since we are interested in the behaviour of the information-acquisition game as k, and hence o^k, goes to infinity, it will turn out to be useful to consider the asymptotic behaviour of cf as $o^k \to \infty$ as a function of (c_1, c_2, c_3). Indeed, the rest of our approach can be seen as the result of an attempt to make this idea rigorous. As a starting point, we define the relative-makeup vector for all test-outcome sequences.

Definition 4.12. Given a test-outcome sequence S and truth assignment A, the *A-trace* of S, denoted $\mathrm{Tr}_A(S)$, is the vector $\mathrm{Tr}_A(S) = (n^+_{S,A,1}/|S|, \ldots, n^+_{S,A,n}/|S|)$. □

Example 4.13. Consider the sequence of test outcomes

$$S_1 = (v_1 \approx T, v_2 \approx T, v_1 \approx T, v_1 \approx T, v_1 \approx F).$$

S_1 has three instances of $v_1 \approx T$, one instance of $v_1 \approx F$ and one instance of $v_2 \approx T$. So the $\{v_1 \mapsto T, v_2 \mapsto T\}$-trace of S_1 is $(\frac{3}{5}, \frac{1}{5})$; the $\{v_1 \mapsto F, v_2 \mapsto T\}$-trace of S_1 is $(\frac{1}{5}, \frac{1}{5})$. If the last test is $v_2 \approx F$ rather than $v_1 \approx F$, giving us the test-outcome sequence

$$S_2 = (v_1 \approx T, v_2 \approx T, v_1 \approx T, v_1 \approx T, v_2 \approx F),$$

then the $\{v_1 \mapsto T, v_2 \mapsto T\}$-trace of S_2 is also $(\frac{3}{5}, \frac{1}{5})$. The sequence

$$S_3 = (v_1 \approx T, v_2 \approx F, v_1 \approx T, v_1 \approx T, v_1 \approx T)$$

has four instances of $v_1 \approx T$ and one of $v_2 \approx F$, so the $\{v_1 \mapsto T, v_2 \mapsto F\}$-trace of S_3 is $(\frac{4}{5}, \frac{1}{5})$. □

It may seem that counting only the test outcomes that agree with A results in an unacceptable loss of information: indeed, as the example above illustrates, the A-traces of the two distinct test-outcome sequences S_1 and S_2 can be the same, even though S_1 and S_2 will in general lead to different posterior probabilities of a formula being true. However, it turns out that if a test-outcome sequence is optimal, there must be some assignment A for which, in a sense, we do not lose information by taking the A-trace.

Definition 4.14. The test-outcome sequence S and the assignment A are *compatible* if all test outcomes in S are compatible with A: that is, S contains no observations of the form $v_i \approx \neg A(v_i)$. □

Lemma 4.15. *Every optimal test-outcome sequence S is compatible with some assignment A.*

Proof. Immediate from Lemma B.2. □

We can now define a counterpart to the earlier definition of a characteristic fraction that uses only the information that is given by an A-trace.

Definition 4.16. 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

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

□

As expected, the quantities $\mathrm{cf}(\varphi, S)$ and $\mathrm{cf}_A(\varphi, \vec{c}, k)$ are closely related. The following lemma makes precise when we do not lose information by considering the appropriate A-trace.

Lemma 4.17. *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 show that the optimal test-outcome sequences for φ, that is, sequences S that maximise $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - \frac{1}{2}|$, satisfy a particular property. By Lemma 4.10, the optimal test-outcome sequences either minimise $\mathrm{cf}(\varphi, S)$ or $\mathrm{cf}(\neg\varphi, S) = 1/\mathrm{cf}(\varphi, S)$. By Lemma 4.15, there must be some assignment A that S is compatible with; by Lemma 4.17, we can equivalently minimise $\mathrm{cf}_A(\varphi, S)$ or $\mathrm{cf}_A(\neg\varphi, S)$ for a truth assignment A compatible with S. In Appendix B we show that if S is sufficiently long and compatible with A and $\varphi(A) = T$, then we must have $\Pr_{D,\vec{\alpha}}(\varphi|S) \geq \Pr_{D,\vec{\alpha}}(\neg\varphi|S)$, while if $\varphi(A) = F$, the opposite inequality must hold (see Lemma B.1). So we need to minimise $\mathrm{cf}_A(\varphi, S)$ if $\varphi(A) = T$ and to minimise $\mathrm{cf}_A(\neg\varphi, S)$ if $\varphi(A) = F$. It suffices to find a sequence S and a truth assignment A that is compatible with S for which the appropriate cf_A is minimal.

The plan. What does this view actually gain us over the naive undertaking to identify the optimal test-outcome sequences directly? Recall that our definition of rational inattention depends on the asymptotic behaviour of information-acquisition games as the number k of tests goes to infinity. Even without the exponential dependency of the search space on k, it is not clear how to analyze large games.

Here, the machinery of A-traces, which do not depend on k at all, comes in helpful. As part of the statement of Theorem 4.23, we provide an A-trace analogue (that is thus independent of k) of the criterion that Proposition 4.7 gives for rational inattention in test-outcome sequences. In the course of the proof of this theorem in the appendix, we will see that, with certain caveats, we can say enough about the A-traces of all optimal test-outcome sequences to be able to check whether they satisfy the analogous criterion. Roughly speaking, the space of A-traces represents something like a continuous generalisation of (a subspace of) test-outcome sequences. We show that inside this space, there is a convex polytope of what can be viewed as "optimal A-traces" (cf. Lemma B.7). As a convex polytope, it is amenable to well-known and well-behaved optimisation techniques. Moreover, as the number k of tests grows, the A-traces of actual optimal test-outcome sequences get progressively closer to points in this polytope. It follows that if all the points in the polytope of "optimal A-traces" satisfy the criterion for rational inattention, then all the A-traces of actual optimal test-outcome sequences are a negligible function away from satisfying them.

We present the definition of this convex polytope and the formal statement of the A-trace analogue criterion here, but relegate the technical details of the rest of the proof to Appendix B.

Assumption: We assume for ease of exposition in the remainder of the paper that the measurement accuracy 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 require us to parameterize RI by the measurement accuracy; the formulae that exhibit $(0.1, 0.1)$-RI might not be the same as those that exhibit $(0.1, 0.3)$-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 compatible 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 $\mathrm{cf}_A(S)$, which is due to some assignment B for which $\varphi(B) = F$, will in general be smaller than $1 \cdot |S|$, and depends on the structure of φ. On the other hand, if $\varphi(A) = F$, we want to minimise the characteristic fraction for $\neg\varphi$, for which the sets of satisfying and falsifying assignments are the opposite of those with A. So, in either case, the greatest power

in the numerator of the characteristic fraction we care about will be due to an assignment B for which $\varphi(B) \neq \varphi(A)$. As Lemma 4.19 below shows, we can formalise the appropriate highest power as follows:

Definition 4.18. 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 4.19. *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 all entries in $\mathrm{Tr}_A(S)$ are non-negative, and Lemma 4.17. □

We now show that the search for the max-power can be formulated as a linear program (LP). Note that if R is a compact subset of $\mathbb{R}$, finding a maximal element of the set is equivalent to finding a minimal upper bound for it:

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

Hence, finding the vector $\vec{c}$ with $\sum_i c_i = 1$ and $c_i \geq 0$ that attains the greatest max-power, that is, 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 max-power 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.

Definition 4.20. Given a formula φ and truth assignment A, define the *conflict LP* $L_A(\varphi)$ to be the linear program

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

□

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 minimum 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 (i.e., closed and bounded) set. It is almost immediate from the definitions that the solution to $L_A(\varphi)$ is $\sup_{\vec{c}:\sum_{i=1}^n c_i=1,\ c_i\geq 0} \mathrm{maxp}_{\varphi,A}(\vec{c})$.

We call $L_A(\varphi)$ a *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 4.21. 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 $\inf_{p \in Feas(L)} f(p)$ 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, $\mathrm{OPT}(L) = \{p \in Feas(L) : f(p) = \mathrm{MIN}(L)\}$.

□

It now follows that if $(\vec{d}, m) \in \mathrm{OPT}(L_A(\varphi))$, then $\max_{\varphi,A}(\vec{d}) = m = \mathrm{MIN}(L_A(\varphi))$. Our goal is to show that 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.

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 4.22. For a formula φ, the *minimax power* $\mathrm{MIN}^*(\varphi)$ is 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 $\mathrm{MIN}(L_A(\varphi)) = \mathrm{MIN}^*(\varphi)$. □

With this definition, we can finally state the core theorem of this section.

Theorem 4.23. *If there exists a constant $C > 0$ 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 $v_i \leq_\varphi v_j$, $c_i \geq C$, and $c_j = 0$, then φ exhibits RI.*

Proof. See Appendix B.2. □

As the conflict LP is of polynomial size, we can solve for a point in the solution polytope in polynomial time. Some additional subtleties are involved in making sure that the criteria we imposed on the coordinates, which represent the A-trace counterpart of the criteria of Proposition 4.7 for optimal test-outcome sequences, are satisfied for all such points. In Section 4.3, we will see that we can do this by solving a polynomial number of LPs derived from the conflict LP. Thus, we obtain a polynomial-time (in 2^n, where n is the number of variables) algorithm for evaluating a sufficient criterion for a formula to exhibit rational inattention.

4.3 Using LPs for nonconvex properties

Let $P_C(\vec{c}, \varphi)$ denote the property that $\vec{c}$ has entries c_i and c_j such that $v_i \leq_\varphi v_j$, $c_i \geq C$, and $c_j = 0$. By Theorem 4.23, a condition sufficient to guarantee that a formula φ exhibits RI is that there exists a C such that $P_C(\vec{c}, \varphi)$ holds for all $(\vec{c}, m) \in \mathrm{OPT}(L_A(\varphi))$. To compute how many formulae exhibit RI, we want an efficient algorithm to check whether this condition holds. Since the quantification over C is existential, and if $C < C'$, then $P_{C'}(\vec{c}, \varphi)$ implies $P_C(\vec{c}, \varphi)$, all choices of C makes P_C a sufficient condition for RI, with smaller C giving rise to stronger conditions (i.e., ones that hold for more formulae that exhibit RI). In practice, we found no difference for the formulae that we tested between taking C to be 10^{-4}, 10^{-5}, or 10^{-6}, and hence arbitrarily chose $C = 10^{-5}$ for our computations.

LPs (such as $L_A(\varphi)$) are known to be solvable in polynomial time (see, e.g., [4]). However, rather than finding a description of the entire solution polytope $\mathrm{OPT}(L_A(\varphi))$, standard linear programming algorithms such as that of Karmarkar [4] compute only a single point inside the polytope. Since we are interested in whether *all* points in the polytope satisfy P_C, we have to do some additional work before we can leverage standard LP solvers.

In general, to establish whether all points in $\mathrm{OPT}(L)$ for a minimising LP L satisfy a property P, we can compare the objective values attained at feasible points for which P is true to those attained at points for which P is false. If some point where P is true attains a smaller value than all points where P is false, then no $\neg P$ point can be optimal, and so P is true for all optimal points; likewise, if some point where P is false attains a smaller value than all points where P is true, then P is false for all optimal points. If neither relationship holds, then points in $\mathrm{OPT}(L)$ can satisfy both P and its negation. Of course, it is unclear how we would compare all pairs of points from two infinite sets in general. However, if we know the minimum objective value m^+ across all feasible points that satisfy P, then the first property above can be simplified to say: is m^+ alone smaller than the objective at any point where P is false? (If the minimum exists, it is by definition attained at some point. Conversely, if the objective at some point that satisfies P is smaller than the objective at all points that don't satisfy P, then m^+ must also be smaller than all these points by transitivity of $<$.)

It would be convenient if we could indeed determine such an m^+ for P_C, for instance by solving another LP. However, the subset of feasible points that satisfy P_C may not actually be a convex polytope. Indeed, a priori it may not even be well-defined, as the minimum of a linear function may not be attained on a set that is not closed. In fact, the property that we care about, that is, the existence of indices i and j such that $v_i \leq_\varphi v_j$, $c_i \geq C$, 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, and $C_1 = 0.15$, 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 $c_i \geq C$ and $c_j = 0$ can be imposed easily on a feasible solution 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 closed 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 4.24. 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, $\bigcup L^+_{A,i,j}(\varphi, C) = \{\vec{p} \in \mathit{Feas}(L_A(\varphi)) : P_C(\vec{p}, \varphi)\}$, so $\min_{i,j} \mathrm{MIN}(L^+_{A,i,j}) = m^+$. It would be convenient if we could similarly determine a minimum m^- for all points where $\neg P_C$, and thereby get rid of the quantification over those points as well and simply compare m^+ to m^-. The negation of P_C is equivalent to

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

if a variable v_j is ignored, then every variable v_i that is no more important than v_j is also "pretty much ignored", that is, not even given a C-fraction of tests for the cutoff constant C we picked. Analogously to before, we can now define a collection of convex polytopes whose union is the set of points on which P_C does not hold. For each polytope, we just fix the (not necessarily single) most important variable v_j that is ignored. Let

$$S^-_{A,j} = \{(c_1, \ldots, c_n, m) \in \mathit{Feas}(L_A(\varphi)) : c_i < C \; \forall i : v_i \leq_\varphi v_j, c_i > 0 \; \forall i : v_i \not\leq_\varphi v_j\}.$$

Observe then that, indeed, $\bigcup_j S^-_{A,j}(\varphi, C) \supseteq \{\vec{p} \in \mathit{Feas}(L_A(\varphi)) : \neg P_C(\vec{p}, \varphi)\}$. Unfortunately, the definition of each $S^-_{A,j}$ involves some strict inequalities, so the sets are not closed. Therefore, we can't use standard LP techniques to find m^-, and indeed, the minimum might not even be attained on the set.

While we may not be able to minimise linear functions over non-closed convex polytopes, we can make use of a standard trick to determine whether such a polytope is at least nonempty. This turns out to be sufficient for our purposes.

Proposition 4.25. *We can decide, in time polynomial in the number of variables and the number of bits required to describe the inequalities, whether a set that is defined by linear inequalities (strict or non-strict) is empty.*

Proof. Take the inequalities defining a set U to be $f_1(\vec{x}) \leq c_1, \ldots, f_n(\vec{x}) \leq c_n$ and $g_1(\vec{x}) < d_1, \ldots, g_m(\vec{x}) < d_m$. Then the LP

$$\begin{array}{lrl} \text{minimise} & s & \\ \text{subject to} & f_1(\vec{x}) & \leq c_1 \\ & & \vdots \\ & f_n(\vec{x}) & \leq c_n \\ & g_1(\vec{x}) & \leq c_1 + s \\ & & \vdots \\ & g_m(\vec{x}) & \leq c_m + s \end{array}$$

has a solution $s^* < 0$ iff U is nonempty: If the LP has a solution $s^* < 0$ then the solution point $\vec{x}^*$ also satisfies the inequalities defining U; conversely, all points $x \in U$ must satisfy all inequalities, so the non-strict inequalities in the LP are immediately satisfied. For the strict ones, there exist $s_1, \ldots, s_m < 0$ such that $g_1(x) = c_1 + s_1 < c_1, \ldots, g_m(x) = c_m + s_m < c_m$. Hence, taking $s = \max_i s_i < 0$, (x, s) satisfies the corresponding non-strict LP inequalities as well, and the solution $s^* \leq s < 0$.

This LP has one more variable than the original set of inequalities, and clearly can be described using at most a polynomially greater number of bits than the original under any reasonable encoding. The result follows by using Karmarkar's algorithm [4]. □

As we noted earlier, we can establish that P_C is true if $m^+ < m$ for all m such that $(x, m) \in \mathit{Feas}(L_A(\varphi))$ and $\neg P_C((x, m), \varphi)$. This is true iff *no* point that satisfies $\neg P_C$ attains a value $m \leq m^+$; and each such point must come from at least one of the $S^-_{A,j}$. In other words, defining

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

we require all the sets $T^-_{A,j}$ to be empty. By the proposition above, we can decide this in polynomial time.

Theorem 4.26. *Fix $C > 0$ and set $m^+_C = \min_{A,i,j} \mathrm{MIN}(L^+_{A,i,j}(\varphi, C))$. If $T^-_{A,j}(\varphi, C, m^+_C) = \varnothing$ for all A and j, then φ exhibits rational inattention.*

Proof. As explained above, the sets $T^-_{A,j}$ being empty implies that there is no point satisfying $\neg P_C$ and attaining a max-power of $m \leq m^+_C$. At the same time, m^+_C being the minimum over all inattentive LPs means that the minimum of m over points satisfying P_C in any L_A is m^+_C. Therefore, m^+_C is the minimax power, and all solution points of relevant LPs satisfy P_C. Hence, by Theorem 4.23, φ exhibits RI. □

Corollary 4.27. *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.*

4.4 Rational inattention is widespread

Using this approach, we were able to exhaustively test all formulae that involve at most four 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 relevant 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 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 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 two columns in the row labeled n is 2^{2^n}.) As the results show, at least 15% of formulae exhibit RI.

n	exhibit RI	unknown
1	0	4
2	8	8
3	40	216
4	9952	55584

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	unknown
5	585	3415
6	506	3494
7	293	3707
8	234	3766
9	194	3806

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 real-life interest 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 histories 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 [3, 7, 9, 13]. 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* $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi)$ 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 $k = 0$. So $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi) = 0$.
(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 tests v_1 three times and guesses the appropriate truth value iff all three tests agree has positive expected payoff. It follows that $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi) = 3$.
(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, and $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi) = 7$.
(4) As Example 2.1 shows, $\mathrm{cpl}_{\mathrm{D,q},\alpha}(v_1 \vee v_2) = 2$, and likewise for all other two-variable conjunctions and disjunctions by symmetry.

It is not hard to see that T and F always have complexity 0, while disjunctions and conjunctions have low complexity. Peraps somewhat counterintuitively, the disjunction $v_1 \vee v_2$ has *lower* complexity than v_1; moreover, the larger k is, the lower the complexity of $v_1 \vee \ldots \vee v_k$. This can be intuitively justified by noting that this measure of complexity captures how much work it takes to attain *certainty* about the truth value of a formula. Longer disjunctions are progressively more likely to be true *a priori*; moreover, any evidence (in the form of test outcomes) that a given disjunction of k variables is true is at least as compelling evidence that an extension of this disjunction to $k + 1$ variables is. Therefore, even though the formula looks more complex, the case for it being true actually becomes easier to make.

We can also 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 a formula is antisymmetric in all variables iff it is equivalent to an XOR or a negation of one. Given a formula φ, the *antisymmetrisation* φ_v *of* φ *along* v is

is the formula

$$\varphi_v = (v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \neg\varphi|_{v=\mathrm{T}}),$$

where $\varphi|_{v=x}$ denotes the formula that results from replacing all occurrences of v in φ by x. It is easy to chek that φ_v is indeed antisymmetric in v. 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. □

The following example illustrates how the antisymmetrisation process affects test complexity.

Example 5.3. Consider the formula $\varphi = (a \wedge c) \vee (\neg a \wedge b)$. In order to better separate the complexity values, we increase the threshold q slightly, from 15/34 to 31/68, which can be achieved by setting *bad* and *good* payoffs of −65 and 3, respectively. We then consider the $(D_u, 31/68, \vec{\alpha})$-test complexity of this formula, as well as different successive antisymmetrisations.

Formula	Equivalent Formula	$\mathrm{cpl}_{\mathrm{D,q},\alpha}$
φ	$(a \wedge c) \vee (\neg a \wedge b)$	6
φ_a	$(a \wedge c) \vee (\neg a \wedge \neg c)$	8
$(\varphi_a)_c$	$(a \wedge c) \vee (\neg a \wedge \neg c)$	8
$((\varphi_a)_c)_b$	$a \oplus b \oplus c$	12
φ_b	$(b \wedge \neg a) \vee (b \wedge c) \vee (\neg b \wedge a \wedge \neg c)$	7
$(\varphi_b)_a$	$a \oplus b \oplus c$	12

□

Theorem 5.2 does not rule out the possibility that there are formulae other than those equivalent to the n-variable XOR or its negation that maximise test complexity. We conjecture that this is not the case except when $q = 0$; this conjecture is supported by experiments we've done with formulas that have fewer than eight variables.

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. This suggests that there might be 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 [3, 7]. (This may be related to the fact that such formulae are linearly separable, and hence learnable by support vector machines [12].) 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.

In computer science and electrical engineering, *binary decision diagrams* (BDDs) [6] are used as a compact representation of Boolean functions. BDDs resemble our notion of a testing strategy, although they do not usually come with a notion of testing error or acceptable error margins on the output (guess). Conversely, we could view testing strategies as a generalisation of BDDs, in which we could "accidentally" take the wrong branch (testing noise), a given variable can occur multiple times, leaf nodes can also be labelled "no guess", and the notion of correctness of a BDD for a formula is relaxed to require only that the output be correct with a certain probability. The *expected decision depth* problem of Ron, Rosenfeld, and Vadhan [8] asks how many nodes of a BDD need to be visited in expectation in order to evaluate a Boolean formula; this can also be seen as a measure of complexity. In our setting, an optimal strategy for the "noiseless" information-acquisition game ($\alpha = 1/2$, $-\infty$ payoff for guessing wrong) exactly corresponds to a BDD for the formula; asking about the depth of the BDD amounts to asking about whether the strategy uses more than a given number of tests.

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 has only 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 nine variables. We argue that this certainly covers most concepts that naturally arise in human discourse. Though it is certainly the case that many propositions (e.g., the outcome of elections) depend on many more variables, human speech and reasoning, for reasons of utterance economy if nothing else, usually involves reducing these to simpler compound propositions (such as the preferences of particular blocks of voters). 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).

In our model, we allowed agents to choose to make no guess for a payoff of 0. We could have removed this option, and instead required them to make a guess. We found this setting to be less amenable to analysis, although there seem to be analogues to our results. For instance, as in our introductory example, it is still rational to keep testing the same variable in a disjunction with a probability that is bounded away from zero, no matter how many tests are allowed. However, since giving up is no longer an option, there is also a probability, bounded away from both 0 and 1, that all variables have to be tested (namely when the formula appears to be false, and hence it must be ascertained that all variables are). The definition of test complexity makes sense in the alternative setting as well, though the values it takes change; we conjecture that the theorem about XOR being hardest can be adapted with few changes.

A CALCULATIONS FOR EXAMPLE 2.1

In this section, we fill in the details of the calculations for Example 2.1. We abuse notation by also viewing formulas, assignments, and test-outcome sequences as events in (i.e., subsets of) the space of histories of the game described in Section 2. Specifically,

- we identify a truth asignment A to the n variables in the game with the event consisting of all histories where A is the assignment chosen by nature;
- we identify the formula φ with the event consisting of all histories where φ is true under the assignment A chosen by nature; thus, φ is the disjoint union of all events A such that $\varphi(A) = T$;
- we identify a test-outcome sequence $S = (v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k)$ of length k with the event consisting of all histories where at least k tests are performed, and the outcomes of the first k are described by S.

Observe that with the "good" payoff being +1 and the "bad" payoff being −16, the expected payoff from guessing that the formula is true after observing S is $\Pr_{D,\vec{\alpha}}(\varphi \mid S) \cdot 1 - \Pr_{D,\vec{\alpha}}(\neg\varphi \mid S) \cdot 16$, so it is greater than 0 if and only if $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > 16/17$.

Henceforth, for brevity, let $A_{bb'}$ ($b, b' \in \{\mathrm{T}, \mathrm{F}\}$) refer to the assignment $\{v_1 \mapsto b, v_2 \mapsto b'\}$. By assumption, all test outcomes are independent conditional on a fixed assignment. Suppose first the player tests the same variable twice, say v_1. Then for the "ideal" test outcome sequence $S = (v_1 \approx \mathrm{T}, v_1 \approx \mathrm{T})$, the conditional probability of S given that nature picked A is $(3/4) \cdot (3/4)$ if $A(v_1) = \mathrm{T}$, and $(1/4) \cdot (1/4)$ otherwise. It follows that

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(v_1 \vee v_2 \mid S) \\
= \; & \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}} \mid S) + \Pr_{D,\vec{\alpha}}(A_{\mathrm{TF}} \mid S) + \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}} \mid S) \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(S|A_{\mathrm{TT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}}) + \ldots + \Pr_{D,\vec{\alpha}}(S|A_{\mathrm{FT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}})}{\Pr_{D,\vec{\alpha}}(S)} \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(S|A_{\mathrm{TT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}}) + \ldots + \Pr_{D,\vec{\alpha}}(S|A_{\mathrm{FT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}})}{\sum_A \Pr_{D,\vec{\alpha}}(S|A) \Pr_{D,\vec{\alpha}}(A)} \\
= \; & \frac{((3/4)\cdot(3/4) + (3/4)\cdot(3/4) + (1/4)\cdot(1/4))\cdot(1/4)}{((3/4)\cdot(3/4) + (3/4)\cdot(3/4) + (1/4)\cdot(1/4) + (1/4)\cdot(1/4))\cdot(1/4)} \\
= \; & \frac{(19/16)\cdot(1/4)}{(20/16)\cdot(1/4)} \\
= \; & 19/20 > 16/17.
\end{aligned}$$

Thus, the agent will guess true after observing S, and get a positive expected payoff (since S will be observed with positive probability) as a consequence of testing v_1 twice. Symmetrically, testing v_2 twice gives a positive expected payoff.

On the other hand, suppose the player tests two different variables. The best case would be to get $S = (v_1 \approx \mathrm{T}, v_2 \approx \mathrm{T})$. As before, the probability of S conditioned on some assignment is the product of the probabilities for each of its entries being observed; for instance, $\Pr_{D,\vec{\alpha}}(S \mid A_{\mathrm{TF}}) = (3/4) \cdot (1/4)$. So we get

$$\begin{aligned}
& \Pr_{D,\vec{\alpha}}(v_1 \vee v_2 \mid S) \\
= \; & \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}} \mid S) + \Pr_{D,\vec{\alpha}}(A_{\mathrm{TF}} \mid S) + \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}} \mid S) \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(S|A_{\mathrm{TT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}}) + \ldots + \Pr_{D,\vec{\alpha}}(S|A_{\mathrm{FT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}})}{\Pr_{D,\vec{\alpha}}(S)} \\
= \; & \frac{\Pr_{D,\vec{\alpha}}(S|A_{\mathrm{TT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{TT}}) + \ldots + \Pr_{D,\vec{\alpha}}(S|A_{\mathrm{FT}}) \Pr_{D,\vec{\alpha}}(A_{\mathrm{FT}})}{\sum_A \Pr_{D,\vec{\alpha}}(S|A) \Pr_{D,\vec{\alpha}}(A)} \\
= \; & \frac{((3/4)\cdot(3/4) + (3/4)\cdot(1/4) + (1/4)\cdot(3/4))\cdot(1/4)}{((3/4)\cdot(3/4) + (3/4)\cdot(1/4) + (1/4)\cdot(3/4) + (1/4)\cdot(1/4))\cdot(1/4)} \\
= \; & \frac{(15/16)\cdot(1/4)}{(16/16)\cdot(1/4)} \\
= \; & 15/16 < 16/17.
\end{aligned}$$

An analogous calculation shows that if either of the tests comes out false, the conditional probability is even lower. Thus, after testing different variables, the agent will not make a guess, no matter what the outcome, and so has an expected payoff of 0.

So, indeed, measuring the same variable twice is strictly better than measuring each of them once.

B QUANTIFYING RATIONAL INATTENTION

Our goal is to show that a large proportion of Boolean formulae exhibit RI. To this end, we would like a method to establish that a particular formula exhibits RI that is sufficiently efficient that we can run it on all formulae of a given size, or at least a statistically significant sample. Throughout this section, we focus on some arbitrary but fixed formula φ in n variables $v_1, \ldots, v_n$. Proposition 4.7 gives a sufficient criterion for φ to exhibit RI 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 $\Pr_{D,\vec{\alpha}}(\varphi \mid S)$ (Lemma 4.10) (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 $v_i \approx b$ conditional 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 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 $0 \le d_j, f_j \le 1$. 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 $eo^{|S|}$. We call the factor d_i before $|S|$ in the exponent of the leading term of the numerator the *max-power* (Definition 4.18) 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 minimal by solving a linear program (Definition 4.20).

B.1 Properties of test-outcome sequences

In this subsection, we present some preliminary results that will prove useful in quantifying RI. We start with a lemma that gives a straightforward way of calculating $\Pr_{D,\vec{\alpha}}(A \mid S)$ for an assignment A and a test-outcome sequence S. The lemma also shows that, as the notation suggests, the probability is independent of the strategy σ.

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 $S = (v_{i_1} \approx b_1, \ldots, v_{i_k} \approx b_k)$. We want to compute $\Pr_{D,\vec{\alpha},\sigma}(S \mid A')$ for an arbitrary truth assignment A'. Recall that a strategy σ is a function from test-outcome sequences to a distribution over actions. We write $\sigma_S(\text{test } v)$ to denote the probability that σ tests v given test-outcome sequence S and use $(\,)$ for the empty sequence; more generally, we denote by $\mathrm{test}_j(v)$ the event that the jth variable chosen was v. Then

$$\begin{aligned} \Pr_{D,\vec{\alpha},\sigma}(S \mid A') = \; & \sigma_{(\,)}(\mathrm{test}_1(v_{i_1})) \Pr_{D,\vec{\alpha},\sigma}((v_{i_1} \approx b_1) \mid \mathrm{test}_1(v_{i_1}), A') \ldots \\ & \sigma_{(v_{i_1} \approx b_1, \ldots, v_{i_{k-1}} \approx b_{k-1})}(\mathrm{test}_k(v_{i_k})) \Pr_{D,\vec{\alpha},\sigma}((v_{i_k} \approx b_k) \mid \mathrm{test}_k(v_{i_k}), A'). \end{aligned}$$

Here, we were able to write $\Pr_{D,\vec{\alpha},\sigma}((v_{i_j} \approx b_j) \mid \mathrm{test}_j(v_{i_j}), A')$ without conditioning on the entire test-outcome sequence up to $v_{i_{j-1}}$ because by the definition of the information-acquisition game,

all observations are independent of each other conditioned on the assignment A'. Observe that the terms $\sigma_{(\,)}(\text{test}_1(v_{i_1})), \ldots, \sigma_{(v_{i_1} \approx b_1, \ldots, v_{i_{k-1}} \approx b_{k-1})}(\text{test}_k(v_{i_k}))$ 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, probabilities of the form $\Pr_{D,\vec{\alpha},\sigma}((v_{i_j} \approx b_j) \mid \text{test}_j(v_{i_j}), A')$ do not depend on the strategy σ, so we can drop it from the subscript of $\Pr_{D,\vec{\alpha},\sigma}$; the probability also does not depend on the results of earlier tests (since, by assumption, test outcomes are independent, conditional on the truth assignment). Thus, it follows that

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

Next, we multiply both the numerator and the denominator of this fraction by $\prod_{j=1}^{k} \frac{1}{1/2-\alpha_{i_j}}$. This amounts to multiplying the jth term in each product by $\frac{1}{1/2-\alpha_{i_j}}$. Thus, in the numerator, if $b_j = A(v_{i_j})$, then the jth term in the product equals 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 the 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 an intuitive 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 any 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 $n_1 > n_2$. 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 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 all truth assignments A', $r_{D,\vec{\alpha}}(A', S_k) = r_{D,\vec{\alpha}}(A', S_k^-)$ if $A'(v_i) \neq b$, and $r_{D,\vec{\alpha}}(A', S_k) = o_i r_{D,\vec{\alpha}}(A', S_k^-)$ 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, we can take

$$x_1 = \sum_{\{A:\varphi(A)=T, A(v_i)=b\}} r_{D,\vec{\alpha}}(S_k, A), \qquad x_2 = \sum_{\{A:\varphi(A)=T, A(v_i)\neq b\}} r_{D,\vec{\alpha}}(S_k, A),$$
$$y_1 = \sum_{\{A:A(v_i)=b\}} r_{D,\vec{\alpha}}(S_k, A), \text{ and} \qquad y_2 = \sum_{\{A:A(v_i)\neq b\}} r_{D,\vec{\alpha}}(S_k, A).$$

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

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

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. □

We conclude with a proof of the lemma that relates the ordering of probabilities to that of characteristic fractions.

Proof. (Lemma 4.10) Since, for $x, y > 0$, we have that $x > y$ iff $(1/x) < (1/y)$, it follows from Lemma B.1 that $\Pr_{D,\vec{\alpha}}(\varphi \mid S) > \Pr_{D,\vec{\alpha}}(\varphi \mid S')$ iff

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

which is true iff

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

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. □

B.2 Proof of Theorem 4.23

To prove Theorem 4.23, we 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.3. Given a constant c and negligible function f, a test-outcome sequence S is (f, c, φ)-*good* if there exist variables v_i and v_j such that $v_i \geq_\varphi v_j$, S contains at least $c|S|$ tests of v_j, and S contains at most $f(|S|)$ tests of v_i. S is (f, c, φ)-*bad* if it is not (f, c, φ)-good. □

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

Lemma B.4. *If, for all negligible functions f and constants $c > 0$, there exists an (f, c, φ)-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, φ)-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 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, φ)-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, φ)-bad test-outcome sequences that are optimal for φ, D and $\vec{\alpha}$. We can assume $\mathcal{S}_{f,c}$ is nonempty; otherwise the claim trivially holds. 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, by way 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.3 (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. Indeed, by choice of $\hat{k}$, S is (f, c, φ)-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

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

We further facilitate working with different-length vectors by using $(\vec{v}, \vec{w})$ to denote vector concatenation, so $(\vec{v}, \vec{w})$ denotes the vector $(v_1, \ldots, v_n, w_1, \ldots, w_m)$.

The following fact about LPs will prove useful.

Lemma B.5. *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 $\vec{p} \in Feas(L)$, either $\vec{p}$ is within ϵ of a solution point, that is,*

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

or $f(\vec{p})$ is more than ϵ' away from the optimum, that is,

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

Proof. We will argue by contradiction. Suppose that the claim does not hold, and let Q be the set of all points in $Feas(L)$ that do not satisfy the first inequality; that is,

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

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 $\epsilon' = \inf_{\vec{q} \in Q}(f(\vec{q}) - \mathrm{MIN}(L))/2$ since then, for every point $\vec{p} \in Feas(L)$, if $f(\vec{p}) - \mathrm{MIN}(L) \leq \epsilon'$, then $\vec{p} \notin Q$, and hence by definition of Q, $\vec{p}$ must be 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 $(\vec{q}'_i)_{i=1}^{\infty}$. Write $\vec{q}^*$ for $\lim_{i \to \infty} \vec{q}'_i$. 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 contradiction. □

We have seen how to distill the information in a test-outcome sequence for a formula in n variables into a vector in $\mathbb{R}^n$ by taking A-traces. The following lemma is to be understood as an approximate converse of this process: given a vector in $\mathbb{R}^n$, 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.6. *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 $\vec{d} = (0.3, 0.7)$ 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 $e \le n$). Hence, for each assignment B,

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

Recalling the definition of the max-power for a 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 the traces of optimal test-outcome sequences. While the traces of optimal sequences may not be in $\mathrm{OPT}(L_A(\varphi))$, they must get arbitrarily close to it as the length of the sequence gets larger.

Lemma B.7. *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 $(\vec{d}, m) \in \mathrm{OPT}(L_A(\varphi))$, that is,*

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

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

$$\forall(\vec{d}, m) \in \mathrm{OPT}(L_A(\varphi)).\ \|\mathrm{Tr}_A(S) - \vec{d}\|_1 \geq \epsilon, \tag{5}$$

S is 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 $k_{\epsilon_n} < n \leq k_{\epsilon_{n+1}}$.

Fix $\epsilon > 0$ and A. Choose an arbitrary test-outcome sequence S compatible with A satisfying (5). Without loss of generality, we can assume that $\varphi(A) = T$. (If $\varphi(A) = F$, then the lemma follows from applying the argument below to $\neg\varphi$ and the observation that sequences are optimal for φ iff they are optimal for $\neg\varphi$.) The feasible set of the LP $L_A(\varphi)$ is compact by construction. Therefore, we can invoke Lemma B.5 to obtain a constant $\epsilon_A > 0$, depending on ϵ and the LP (and hence A), such that for all feasible points $p = (\vec{c}, m) \in \mathit{Feas}(L_A(\varphi))$, either $\|\vec{c} - \vec{d}\|_1 < \epsilon$ for some $\vec{d} \in \mathrm{OPT}(L)$, or $|m - \mathrm{MIN}(L_A(\varphi))| > \epsilon_A$. Set

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

(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. By assumption, S satisfies (5). Since appending another entry to a vector can only make its 1-norm greater, we also have $\|(\mathrm{Tr}_A(S), \mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))) - \vec{d}\|_1 \geq \epsilon$ for all $\vec{d} \in \mathrm{OPT}(L_A(\varphi))$. The point $(\mathrm{Tr}_A(S), \mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)))$ is in the feasible set of $L_A(\varphi)$; this contradicts the first option in the disjunction provided by Lemma B.5. Therefore, the second option must be true:

$$|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{MIN}(L_A(\varphi))| \quad > \quad \epsilon_A. \tag{6}$$

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

$$\begin{aligned} & \mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|) \\ = \ & \frac{\sum_{\{B:\varphi(B)=\mathrm{F}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} \mathrm{Tr}_A(S)_i |S|}}{\sum_{\{B:\varphi(B)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} \mathrm{Tr}_A(S)_i |S|}} \\ \geq \ & \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} \qquad \text{[see below]}. \end{aligned} \tag{7}$$

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 $2^n o^{|S|}$.

Fix $(\vec{d}, m) \in \mathrm{OPT}(L_A(\varphi))$. By Lemma B.6, there exists a test-outcome sequence $S_{|S|,\vec{d},A}$ such that $|\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S_{|S|,\vec{d},A})) - \mathrm{maxp}_{\varphi,A}(\vec{d})| < 2n/|S|$. For brevity, set $\vec{d}' = \mathrm{Tr}_A(S_{|S|,\vec{d},A})$. 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 $(\vec{d}, m) \in \mathrm{OPT}(L_A(\varphi))$, we have that $\mathrm{maxp}_{\varphi,A}(\vec{d}) = m = \mathrm{MIN}(L_A(\varphi))$, so $|\mathrm{maxp}_{\varphi,A}(\vec{d}') - \mathrm{MIN}(L_A(\varphi))| < \epsilon_A/2$. Now using (6) 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

$$\begin{aligned} & \mathrm{cf}_A(\varphi, \vec{d}', |S|) \\ = \; & \frac{\sum_{\{B:\varphi(B)=\mathrm{F}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} d'_i|S|}}{\sum_{\{B:\varphi(B)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(B) o^{\sum_{\{v_i:A(v_i)=B(v_i)\}} d'_i|S|}} \\ \leq \; & \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d}')|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}}, \end{aligned} \tag{9}$$

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.

Now observe that

$$\begin{aligned} & \Pr_{D,\vec{\alpha}}(\varphi \mid S_{|S|,\vec{d},A}) > \Pr_{D,\vec{\alpha}}(\varphi \mid S) \\ \text{iff} \quad & \mathrm{cf}(\varphi, S_{|S|,\vec{d},A}) < \mathrm{cf}(\varphi, S) && \text{[by Lemma 4.10]} \\ \text{iff} \quad & \mathrm{cf}_A(\varphi, \vec{d}', |S|) < \mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|) && \text{[by Lemma 4.17]} \\ \text{if} \quad & \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d}')|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}} < \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} && \text{[by (7) and (9)]} \\ \text{iff} \quad & \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}} - \frac{2^n o^{\mathrm{maxp}_{\varphi,A}(\vec{d}')|S|}}{\Pr_{D,\vec{\alpha}}(A) o^{|S|}} > 0 \\ \text{iff} \quad & \frac{\Pr_{D,\vec{\alpha}}(A) \min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S|} - 2^{2n} o^{\mathrm{maxp}_{\varphi,A}(\vec{d}')|S|}}{\Pr_{D,\vec{\alpha}}(A) 2^n o^{|S|}} > 0 \\ \text{iff} \quad & \Pr_{D,\vec{\alpha}}(A) \min_B \Pr_{D,\vec{\alpha}}(B) o^{\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S))|S| - \mathrm{maxp}_{\varphi,A}(\vec{d}')|S|)} - 2^{2n} > 0 \\ \text{iff} \quad & (\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d}'))|S| > \log_o \left(\frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(A) \min_B \Pr_{D,\vec{\alpha}}(B)} \right). \end{aligned} \tag{10}$$

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), $(\mathrm{maxp}_{\varphi,A}(\mathrm{Tr}_A(S)) - \mathrm{maxp}_{\varphi,A}(\vec{d}')) > \epsilon_A/2$. It follows that the last line of (10) is in fact satisfied. Thus 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.8. *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. The proof reuses many of the core ideas of Lemma B.7 in a simplified setting. For contradiction, suppose that A is not relevant, but S is optimal. Let B be an arbitrary relevant assignment. Then

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

We show that we can choose a k_0 such that if $|S| > k_0$, then there is a test-outcome sequence S' of the same length supporting B that is actually better, contradicting the optimality of S.

Indeed, set

$$k_0 = \max \left\{ 4n/\epsilon, \frac{2}{\epsilon} \log_o \left(\frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(B) \min_C \Pr_{D,\vec{\alpha}}(C)} \right) \right\}.$$

Since $\mathrm{Tr}_A(S)$ is a feasible point of L_A, we have $\mathrm{maxp}_{\varphi,A}\mathrm{Tr}_A(S) \geq \mathrm{MIN}(L_A) \geq \mathrm{MIN}(L_B) + \epsilon$. On the other hand, let $(\vec{d}, m) \in \mathrm{OPT}(L_B(\varphi))$ be arbitrary. Since $|S| > 4n/\epsilon$, the B-trace $\vec{d}' = \mathrm{Tr}_B(S_{k,\vec{d},B})$ of the sequence $S_{k,\vec{d},B}$ of Lemma B.6 satisfies

$$|\mathrm{maxp}_{\varphi,A}(\vec{d}') - \mathrm{maxp}_{\varphi,A}(\vec{d})| = |\mathrm{maxp}_{\varphi,A}(\vec{d}') - \mathrm{MIN}(L_B)| < \epsilon/2.$$

So $\max p_{\varphi,A}(\mathrm{Tr}_A(S)) - \max p_{\varphi,A}(\vec{d}') > \epsilon/2$. As in the proof of Lemma B.7, we have

$$\mathrm{cf}_A(\varphi, \mathrm{Tr}_A(S), |S|) \geq \frac{\min_B \Pr_{D,\vec{\alpha}}(B) o^{\max p_{\varphi,A}(\mathrm{Tr}_A(S))|S|}}{2^n o^{|S|}}$$

for S and

$$\mathrm{cf}_B(\varphi, \vec{d}', |S|) \leq \frac{2^n o^{\max p_{\varphi,A}(\vec{d}')|S|}}{\Pr_{D,\vec{\alpha}}(B) o^{|S|}}$$

for the sequence that approximates $\vec{d}$, and hence

$$\begin{array}{ll} & \Pr_{D,\vec{\alpha}}(\varphi \mid S_{|S|,\vec{d},B}) > \Pr_{D,\vec{\alpha}}(\varphi \mid S) \\ \text{iff} & \mathrm{cf}(\varphi, S_{|S|,\vec{d},B}) < \mathrm{cf}(\varphi, S) \\ \text{if} & (\ldots) \\ \text{iff} & (\max p_{\varphi,A}(\mathrm{Tr}_A(S)) - \max p_{\varphi,A}(\vec{d}'))|S| > \log_o \left(\frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(B) \min_C \Pr_{D,\vec{\alpha}}(C)} \right). \end{array}$$

But $|S| > k_0 \geq \frac{2}{\epsilon} \log_o \left(\frac{2^{2n}}{\Pr_{D,\vec{\alpha}}(B) \min_C \Pr_{D,\vec{\alpha}}(C)} \right)$, and hence S is indeed not optimal. □

With these pieces, we can finally prove Theorem 4.23.

Proof (of Theorem 4.23). Suppose, by way of contradiction, that the antecedent of Theorem 4.23 holds, but φ does not exhibit RI. Let δ be the function of Lemma B.7 and let C be the constant that is assumed to exist in the statement of Theorem 4.23. 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 Proposition 4.7, there exists an open-minded product distribution D and accuracy vector α such that for all c, there exists an (f, c)-bad test-outcome sequence optimal for φ, D, and $\vec{\alpha}$. So by Lemma B.4, taking $c = C/2$, there exists an infinite sequence $\{S_k\}$ of $(f, C/2, \varphi)$-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 $C|S_k|/2$ times. (11)

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.8, we can assume that A is relevant, since all but finitely many of the S_k must be sufficiently long.

Let k_1 be sufficiently large that $\delta(k) < C/2$ for all $k > k_1$. By Lemma B.7, for all $k > k_1$, we must have

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

for some solution $(\vec{d}, m)$ 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 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 (11). □

C PROOF OF THEOREM 5.2

We previously took the XOR $v_1 \oplus \ldots \oplus v_n$ of n variables (often denoted $\bigoplus_{i=1}^{n} v_i$) to be true iff an odd number of the variables are true. This characterisation is actually a consequence of the following standard definition in terms of basic Boolean connectives, of which we also note some useful properties (whose proof is left to the reader).

Definition C.1. The *exclusive OR* (XOR) $\varphi_1 \oplus \varphi_2$ is equivalent to the formula $(\varphi_1 \wedge \neg\varphi_2) \vee (\neg\varphi_1 \wedge \varphi_2)$. □

Proposition C.2. *(Properties of XOR)*

(a) XOR is commutative: $\varphi_1 \oplus \varphi_2 \equiv \varphi_2 \oplus \varphi_1$*;*
(b) XOR is associative: $(\varphi_1 \oplus \varphi_2) \oplus \varphi_3 \equiv \varphi_1 \oplus (\varphi_2 \oplus \varphi_3)$*;*
(c) $v_1 \oplus \ldots \oplus v_n$ *is true iff an odd number of the variables* v_i *is;*
(d) $\neg\varphi \equiv T \oplus \varphi$*, so* $\varphi_1 \oplus \neg\varphi_2 \equiv \neg\varphi_1 \oplus \varphi_2 \equiv \neg(\varphi_1 \oplus \varphi_2)$.

As we said in the proof sketch in the main text, our proof uses the idea of antisymmetry.

The notion of antisymmetry has the useful property that φ_v, the antisymmetrisation of φ along v (recall that φ_v was defined as $(v \wedge \varphi|_{v=\mathrm{T}}) \vee (\neg v \wedge \neg\varphi|_{v=\mathrm{T}})$) is antisymmetric in v and, as we now show, also antisymmetric in all other variables v' that φ was antisymmetric in.

Lemma C.3. *If* φ *is antisymmetric in a variable* $v' \neq v$*, then so is* φ_v.

Proof. Suppose that φ is antisymmetric in $v' \neq v$. Then for all truth assignments A, we have

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

Thus, if $A(v) = \mathrm{T}$, then

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

and if $A(v) = \mathrm{F}$, then

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

Thus, no matter what $A(v)$ is, we have $\varphi_v(A[v' \mapsto \mathrm{T}]) = \neg\varphi_v(A[v' \mapsto \mathrm{F}])$, as required. □

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

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

Lemma C.4. *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, by Definition C.1 we have that

$$\varphi \equiv (v_n \wedge \varphi|_{v_n=\mathrm{T}}) \vee (\neg v_n \wedge \neg\varphi|_{v_n=\mathrm{T}}) \equiv v_n \oplus \varphi|_{v_n=\mathrm{T}}. \tag{12}$$

It is easy to see that $\varphi|_{v_n=\mathrm{T}}$ mentions only the variables $v_1, \ldots, v_{n-1}$ and, by Lemma C.3, is antisymmetric in each of them. So by the induction hypothesis, $\varphi|_{v_n=\mathrm{T}}$ is equivalent to either $\bigoplus_{i=1}^{n-1} v_i$ or $\neg(\bigoplus_{i=1}^{n-1} v_i)$, and hence by Proposition C.2(d) and (12), φ is equivalent to either $\bigoplus_{i=1}^{n} v_i$ or $\neg(\bigoplus_{i=1}^{n} v_i)$. □

To complete the proof of Theorem 5.2, we make use of the following two technical lemmas. For the remainder of the proof, we use $v = \mathrm{T}$ and $v = \mathrm{F}$ to denote the events (i.e., the set of histories) where the variable v is true (resp., false). (We earlier denoted these events v and $\neg v$, respectively, but for this proof the $v = b$ notation is more convenient.)

Lemma C.5. *If D is a product distribution and S is a test-outcome sequence, then 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. Given a truth assignment A on $v_1, \ldots, v_n$, let A_i be A restricted to all the variables other than v_i. Since D is a product distribution, $\Pr_{D,\vec{\alpha}}(A) = \Pr_{D,\vec{\alpha}}(A_i) \times \Pr_{D,\vec{\alpha}}(v_i = A(v_i))$.

Note that the truth of $\varphi|_{v_i=b}$ does not depend on the truth value of v_i. Thus, we can pair the truth assignments that make $\varphi|_{v_i=b}$ true into groups of two, that differ only in the truth assignment to v_i. Suppose that the test $v_i \approx \mathrm{T}$ appears in S k_T times and the test $v_i \approx \mathrm{F}$ appears in S k_F times. Using Lemma B.1, we have that

$$\begin{aligned}
\Pr_{D,\vec{\alpha}}(\varphi|_{v_i=b} \mid S) &= \textstyle\sum_{\{A:\, \varphi|_{v_i=b}(A)=T\}} \Pr_{D,\vec{\alpha}}(A \mid S) \\
&= \frac{\sum_{\{A:\, \varphi|_{v_i=b}(A)=T\}} r_{D,\vec{\alpha}}(A,S)}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A',S)} \\
&= \frac{\sum_{\{A:\, \varphi|_{v_i=b}(A)=T\}} r_{D,\vec{\alpha}}(A_i,S)(\Pr_{D,\vec{\alpha}}(v_i=T)o^{k_T} + \Pr_{D,\vec{\alpha}}(v_i=F)o^{k_F})}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A'_i,S)(\Pr_{D,\vec{\alpha}}(v_i=T)o^{k_T} + \Pr_{D,\vec{\alpha}}(v_i=F)o^{k_F})} \qquad (13) \\
&= \frac{\sum_{\{A:\, \varphi|_{v_i=b}(A)=T\}} r_{D,\vec{\alpha}}(A_i,S)}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A'_i,S).}
\end{aligned}$$

Using the same arguments as in (13), we get that

$$\Pr_{D,\vec{\alpha}}(\varphi \wedge v_i = b \mid S) = \frac{\sum_{\{A:\, (\varphi \wedge v_i=b)(A)=T\}} r_{D,\vec{\alpha}}(A_i, S) \Pr_{D,\vec{\alpha}}(v_i = b) o^{k_T}}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S)}$$

and

$$\Pr_{D,\vec{\alpha}}(v_i = b \mid S) = \frac{\sum_{\{A:\, A(v_i)=b\}} r_{D,\vec{\alpha}}(A_i, S) \Pr_{D,\vec{\alpha}}(v_i = b) o^{k_T}}{\sum_{\text{truth assignments } A'} r_{D,\vec{\alpha}}(A', S).}$$

Let $C = \Pr_{D,\vec{\alpha}}(v_i = b \mid S) = \frac{\Pr_{D,\vec{\alpha}}(v_i=b) k_b}{\Pr_{D,\vec{\alpha}}(v_i=T)k_T + \Pr_{D,\vec{\alpha}}(v_i=F)k_F}$ be the probability that $v_i = b$ *after* observing the sequence. Note that

$$\sum_{\{A:\, (\varphi \wedge v_i=b)(A)=T\}} r_{D,\vec{\alpha}}(A_i, S) \quad = \quad \sum_{\{A:\, (\varphi|_{v_i=b} \wedge v_i=b)(A)=T\}} r_{D,\vec{\alpha}}(A_i, S) \quad = \quad C \cdot \sum_{\{A:\varphi|_{v_i=b}(A)=T\}} r_{D,\vec{\alpha}}(A_i, S)$$

and

$$\sum_{\{A:\, A(v_i)=b\}} r_{D,\vec{\alpha}}(A_i, S) \quad = \quad C \cdot \sum_{\text{truth assignments } A} r_{D,\vec{\alpha}}(A_i, S).$$

Since, by Bayes' Rule,

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S, v_i = b) = \frac{\Pr_{D,\vec{\alpha}}(\varphi \wedge v_i = b \mid S)}{\Pr_{D,\vec{\alpha}}(v_i = b \mid S)},$$

simple algebra shows that $\Pr_{D,\vec{\alpha}}(\varphi \mid S, v_i = b) = \Pr_{D,\vec{\alpha}}(\varphi|_{v_i=b} \mid S)$, as desired. □

Lemma C.6. *If, for all test-outcome sequences S, there exists a test-outcome sequence S' such that $|S'| = |S|$ and $|\Pr_{D,\vec{\alpha}}(\varphi \mid S') - 1/2| \geq |\Pr_{D,\vec{\alpha}}(\psi \mid S) - 1/2|$, then $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi) \leq \mathrm{cpl}_{\mathrm{D,q},\alpha}(\psi)$.*

Proof. Suppose that $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\psi) = k$. Then there must be some strategy σ for $G(\psi, D, k, \vec{\alpha}, g, b)$ that has positive expected payoff. There must therefore be some test-outcome sequence S of length k that is observed with positive probability when using σ such that the expected payoff of making the appropriate guess is positive. By Lemma 3.3, $|\Pr_{D,\vec{\alpha}}(\psi \mid S) - 1/2| > q$.

Since $|\Pr_{D,\vec{\alpha}}(\varphi \mid S') - 1/2| \geq |\Pr_{D,\vec{\alpha}}(\psi \mid S) - 1/2|$ by assumption, there must exist a test-outcome sequence S' such $|\Pr_{D,\vec{\alpha}}(\varphi \mid S') - 1/2| > q$. 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 iff 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 $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi)$ is at most k. □

We can now finally prove Theorem 5.2. Note that this is the only part of the derivation that actually depends on the assumption that we are working with the uniform distribution D_u.

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 $\mathrm{cpl}_{\mathrm{D,q},\alpha}\,\varphi \leq \mathrm{cpl}_{\mathrm{D,q},\alpha}\,\varphi_0$. By Lemma C.4, 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 $V(\varphi) > 0$. There there must exist some variable v such that $\varphi|_{v=\mathrm{T}} \neq \neg(\varphi|_{v=\mathrm{F}})$. (Here and below we are viewing formulas as functions on truth assignments, justifying the use of "=" rather than "≡".) Note for future reference that, by construction,

$$\varphi_v|_{v=\mathrm{T}} = \varphi|_{v=\mathrm{T}} \text{ and } \varphi_v|_{v=\mathrm{F}} = \neg\varphi|_{v=\mathrm{T}}. \tag{14}$$

By Lemma C.3, if φ is antisymmetric in a variable $v' \neq v$, then so is φ_v. In addition, φ_v is antisymmetric in v. Thus, $V(\varphi_v) < V(\varphi)$. If we can show $\mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi) \leq \mathrm{cpl}_{\mathrm{D,q},\alpha}(\varphi_v)$, then the result follows from the induction hypothesis. By Lemma C.6, 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 $|\Pr_{D,\vec{\alpha}}(\varphi \mid S) - 1/2| \geq |\Pr_{D,\vec{\alpha}}(\varphi_v \mid S_1) - 1/2|$.

Given an arbitrary test-outcome sequence S_1, let $p = \Pr_{D,\vec{\alpha}}(v = \mathrm{T} \mid S_1)$. Thus,

$$\begin{aligned}
\Pr_{D,\vec{\alpha}}(\varphi_v \mid S_1) &= p \Pr_{D,\vec{\alpha}}(\varphi_v \mid S_1, v = \mathrm{T}) + (1-p) \Pr_{D,\vec{\alpha}}(\varphi_v \mid S_1, v = \mathrm{F}) \\
&= p \Pr_{D,\vec{\alpha}}(\varphi_v|_{v=\mathrm{T}} \mid S_1) + (1-p) \Pr_{D,\vec{\alpha}}(\varphi_v|_{v=\mathrm{F}} \mid S_1) && \text{[by Lemma C.5]} \\
&= p \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p) \Pr_{D,\vec{\alpha}}(\neg\varphi|_{v=\mathrm{T}} \mid S_1) && \text{[by (14)]} \\
&= p \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p)(1 - \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1).
\end{aligned} \tag{15}$$

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. Since $\varphi|_{v=\mathrm{T}}$ does not mention v, $\Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) = \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_2)$ 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)

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S_1) = p \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p) \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{F}} \mid S_1) \tag{16}$$

and, taking $p' = \Pr_{D,\vec{\alpha}}(v = T \mid S_2)$,

$$\begin{aligned}
\Pr_{D,\vec{\alpha}}(\varphi \mid S_2) &= p' \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_2) + (1-p') \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{F}} \mid S_2) \\
&= p' \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + (1-p') \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{F}} \mid S_1).
\end{aligned} \tag{17}$$

We claim that $p = 1 - p'$. Suppose that the test $v \approx \mathrm{T}$ appears in S_1 k_T times and the test $v \approx \mathrm{F}$ appears in S_1 k_F times. Thus, the test $v \approx \mathrm{T}$ appears in S_2 k_F times and the test $v \approx \mathrm{F}$ appears in S_1 k_T times. All other tests appear the same number of times in both sequences. By Lemma B.1, since

the uniform distribution D_u we are using is in particular a product distribution, for $j = 1, 2$, we have that

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

Suppose that v is the ith variable v_i. Let $r_1 = o_i^{k_T}$, let $r_2 = o_i^{k_F}$, let $R_1 = \sum_{\{A:\, A(v_i)=T\}} \prod_{j=1, j\neq i}^{n} o_j^{n^+_{S_1,A,j}}$, and let $R_2 = \sum_{\{A:\, A(v_i)=F\}} \prod_{j=1, j\neq i}^{n} o_j^{n^+_{S_1,A,j}}$. For $j = 1, 2$ we have that

$$\sum_{\{A:\, A(v)=\mathrm{T}\}} \Pr_{D,\vec{\alpha}}(A \mid S_j) = \frac{\sum_{\{A:\, A(v)=\mathrm{T}\}} r_{D,\vec{\alpha}}(A, S_j)}{\sum_{A'} r_{D,\vec{\alpha}}(A', S_j)} = \frac{r_j R_j}{r_1 R_1 + r_2 R_2}$$

We claim that $R_1 = R_2$. Indeed, for any assignment A such that $A(v_i) = T$, let A' be the unique assignment such that $A'(v_i) = F$ and $A'(v_j) = A(v_j)$ for all $j \neq i$. Then each choice of A occurs once in the sum R_1 and never in the sum R_2, the corresponding A' occurs once in R_2 but not R_1. Since we are working with the uniform distribution D_u, the summands for A and A' are equal. So we can conclude that $p = 1 - p'$. Combining this with (17), we get that

$$\Pr_{D,\vec{\alpha}}(\varphi \mid S_2) \quad = \quad (1-p)\Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{T}} \mid S_1) + p \Pr_{D,\vec{\alpha}}(\varphi|_{v=\mathrm{F}} \mid S_1). \tag{18}$$

Let $Q(E) = \Pr_{D,\vec{\alpha}}(E) - \frac{1}{2}$. By adding $-1/2$ on both sides, equations (16) and (18) hold with $\Pr_{D,\vec{\alpha}}$ replaced by Q, while (15) becomes

$$Q(\varphi_v \mid S_1) = pQ(\varphi|_{v=\mathrm{T}} \mid S_1) - (1-p)Q(\varphi|_{v=\mathrm{T}} \mid S_1).$$

We now show that either $|\,\mathrm{Q}(\varphi \mid S_1)| \geq |\,\mathrm{Q}(\varphi_v \mid S_1)|$ or $|\,\mathrm{Q}(\varphi \mid S_2)| \geq |\,\mathrm{Q}(\varphi_v \mid S_1)|$. This suffices to complete the proof.

To simplify notation, let $x = \mathrm{Q}(\varphi|_{v=\mathrm{T}} \mid S_1)$ and let $y = \mathrm{Q}(\varphi|_{v=\mathrm{F}} \mid S_1)$. By (15), (16), and (18), 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 $q > 1/2$. □

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

[1] Y. R. Chen and M. N. Katehakis. 1986. Linear Programming for Finite State Multi-Armed Bandit Problems. *Mathematics of Operations Research* 11, 1 (1986), 180–183.

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

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

[4] N. Karmarkar. 1984. A new polynomial-time algorithm for linear programming. In *Proc. 16th ACM Symposium on Theory of Computing*. 302–311.
[5] J. Lang, P. Liberatore, and P. Marquis. 2003. Propositional independence – Formula-variable independence and forgetting. *Journal of Artificial Intelligence Research* 18 (2003), 391–443.
[6] C. Y. Lee. 1959. Representation of switching circuits by binary-decision programs. *The Bell System Technical Journal* 38 (1959), 985–999.
[7] B. C. Love, D. L. Medin, and T. M. Gureckis. 2004. SUSTAIN: A network model of category learning. *Psychological Review* 111, 2 (4 2004), 309–332.
[8] D. Ron, A. Rosenfeld, and S. Vadhan. 2007. The hardness of the expected decision depth problem. *Inform. Process. Lett.* 101, 3 (2007), 112–118.
[9] R. N. Shepard, C. I. Hovland, and H. M. Jenkins. 1961. Learning and memorization of classifications. *Psychological Monographs: General and Applied* 75, 3 (1961), 1–42.
[10] C. A Sims. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50, 3 (2003), 665–690.
[11] C. Umans. 1999. On the complexity and inapproximability of shortest implicant problems. In *Proc. of Automata, Languages and Programming: 26th International Colloquium (ICALP '99)*. Springer, Berlin, Heidelberg, 687–696.
[12] V. N. Vapnik and A. Y. Lerner. 1963. Recognition of patterns using generalized portraits. *Avtomat. i Telemekh.* 24 (1963), 774–780. Issue 6.
[13] R. Vigo. 2011. Representational information: a new general notion and measure of information. *Information Sciences* 181 (2011), 4847–4859.
[14] M. Wiederholt. 2010. Rational inattention. In *The New Palgrave Dictionary of Economics (online edition)*, L. E. Blume and S. Durlauf (Eds.). Palgrave Macmillan, New York.