Finite-Sample Bounds for Two-Distribution Hypothesis Tests

Cynthia Hom, William Yik, George D. Montañez
{chom, wyik, gmontanez}@hmc.edu
AMISTAD Lab, Department of Computer Science
Harvey Mudd College
Claremont, CA, USA

***Abstract*—With the rapid growth of large language models, big data, and malicious online attacks, it has become increasingly important to have tools for anomaly detection that can distinguish machine from human, fair from unfair, and dangerous from safe. Prior work has shown that two-distribution (specified complexity) hypothesis tests are useful tools for such tasks, aiding in detecting bias in datasets and providing artificial agents with the ability to recognize artifacts that are likely to have been designed by humans and pose a threat. However, existing work on two-distribution hypothesis tests requires exact values for the specification function, which can often be costly or impossible to compute. In this work, we prove novel finite-sample bounds that allow for two-distribution hypothesis tests with only estimates of required quantities, such as specification function values. Significantly, the resulting bounds do not require knowledge of the true distribution, distinguishing them from traditional p-values. We apply our bounds to detect student cheating on multiple-choice tests, as an example where the exact specification function is unknown. We additionally apply our results to detect representational bias in machine-learning datasets and provide artificial agents with intention perception, showing that our results are consistent with prior work despite only requiring a finite sample of the space. Finally, we discuss additional applications and provide guidance for those applying these bounds to their own work.**

***Index Terms*—statistics, hypothesis testing, specified complexity, anomaly detection, machine learning, bias, artificial agents**

I. Introduction

Imagine you're in a survival contest where contestants are thrown into the wilderness and left to defend themselves. The last survivor wins the game, and the other contestants seek to eliminate you from the competition. One contestant is an expert trapper, so you worry your surroundings may contain hidden traps. However, these traps are constructed from branches, leaves, and other natural components, making them difficult to distinguish from random configurations of foliage.

In order to distinguish a natural assemblage of branches from an intentionally constructed trap, you examine the arrangement of pieces in each pile to see how they differ from the typical random bunches of foliage that one might encounter in a forest. Previous work [1] provides a set of hypothesis tests for this task. To determine the likelihood of an assemblage of branches being a trap, the tests take into account both how specific the instance is (in the context of the general space of instances), as well as the probability of encountering the assemblage.

However, consider the following complication–due to an injury, you have diminished sight. Thus, you can't easily determine how many branches are present in this assemblage, nor can you tell their orientations. However, you are able to determine which part of the structure has branches and which has leaves. In this case, it is harder to evaluate the structure in front of you. The structure you imagine may differ from the true form of the object by some margin of error. The hypothesis tests of Montañez [1] require that the exact specificity (degree of structure in relation to the rest of the forest) and probability of encountering the object must be known; however, with injured eyes, you cannot very well determine the exact specificity of the object. Nevertheless, you'd still like to make use of such hypothesis tests.

In this work, we present a set of finite-sample bounds for two-distribution (specified complexity) hypothesis tests [1], thereby enabling their use with sampling procedures rather than exact calculations of specificity. These two-distribution hypothesis tests have been previously applied to several tasks including intention perception in artificial agents [2] and identifying bias in machine learning datasets [3]. In both of these prior applications, the specification function, which gives a numerical value for the degree of structure an object holds, was either computed through manual enumeration or a derived combinatorial formula. While these both led to successful calculations of the specification function and enabled the use of two-distribution hypothesis tests, the function may not always be easily computed, and many practitioners may not wish to spend the time to derive analytical formulae. To enable more widespread adoption of two-distribution hypothesis tests by non-mathematicians, we propose estimation procedures to replace exact computations for specificity, and prove novel measure concentration bounds given such estimates.

We apply our new finite-sample bounds to the same tasks as Hom et al. [2] and Yik et al. [3], obtaining similar bounds without ever calculating the specification function directly, thereby reducing the work required to conduct a two-distribution hypothesis test. We also test our bounds on a scenario where an instructor wishes to identify cheating amongst students on a multiple-choice test. Unlike the previous problems [2], [3], the specification function for this scenario has not yet been

enumerated. Overall, our finite-sample bounds broaden the range of applications for two-distribution hypothesis tests and vastly streamline the testing procedure by requiring only a simple sample to estimate the specification function.

The remainder of this paper is structured as follows. In Section II, we provide relevant mathematical background for two-distribution hypothesis tests [1] and review methods for conducting them which use exact values for the specification function. In Section III, we provide finite-sample bounds on the two-distribution tests which enable their use even when there is only a sampled estimate of the specification function, subject to some error. Section IV details some example scenarios, where we show three distinct applications of the finite-sample tests. Lastly, in Section V, we elaborate on the benefits of our new finite-sample tests, including the reduction of computation required to calculate the specification function. We also discuss applications of these tests such as detecting genetically-modified DNA, distinguishing the work of humans from that of general artificial intelligence, and identifying malicious network activity.[1]

II. Background

This work builds upon the concept of *specified complexity* [4], a measure of both how structurally organized and unlikely an object is. In particular, objects with high specified complexity are both specified (matching a predetermined form) and complex (unlikely to occur under a given probability distribution). A *specified complexity model* consists of a complexity function $p(x)$ which captures how likely an element x is to be selected from a space $\mathcal{X}$, and a specification function $\nu(x)$ which captures how coherent or structured x is. Following Montañez [1], we begin with a few formal definitions.

Definition 1 ($\nu(\mathcal{X})$, [1])**.** *For any integrable nonnegative specification function $\nu : \mathcal{X} \to \mathbb{R}_{\geq 0}$, define $\nu(\mathcal{X})$ as follows:*

$$\nu(\mathcal{X}) \coloneqq \begin{cases} \int_{\mathcal{X}} \nu(x)dx & \text{if continuous,} \\ \sum_{x\in\mathcal{X}} \nu(x) & \text{if discrete,} \\ \int_{\mathcal{X}} d\nu(x) & \text{in general.} \end{cases}$$

Definition 2 (Common Form and Kardis, [1])**.** *For any probability distribution $p(x)$ on space $\mathcal{X}$, any strictly positive scaling constant $r \in \mathbb{R}_{>0}$ and any nonnegative function $\nu : \mathcal{X} \to \mathbb{R}_{\geq 0}$, we define a common form model as*

$$SC(x) \coloneqq -\log_2 r\frac{p(x)}{\nu(x)}$$

with specified complexity kardis $\kappa(x) = r\frac{p(x)}{\nu(x)}$.

Definition 3 (Canonical Specified Complexity Model [1])**.** *Any common form model constrained such that $\nu(\mathcal{X}) \leq r$ is a canonical specified complexity model.*

[1]See source code and experimental notebooks at https://github.com/AMISTAD-lab/finite-sample-bounds.

While there exist many canonical specified complexity models, the Functional Specified Complexity (FSC) model proposed by Montañez [1] (based on functional information [5]) is particularly useful because it works with finite, discrete data and eliminates the need to estimate the normalization factor r. More importantly, FSC gives a more concrete way to compute $\nu(x)$ based on a given function $g : \mathcal{X} \to \mathbb{R}_{\geq 0}$ that increases with increasing degrees of extremity for an observation x. Following [1], we define $M_g(x) = |\{x' \in \mathcal{X} : g(x') \geq g(x)\}|$, which yields the *functional specificity*

$$F_g(x) = \frac{M_g(x)}{|\mathcal{X}|}.$$

With this in mind, we formally define FSC as follows.

Definition 4 (Functional Specified Complexity [1])**.** *For function g, functional specificity $F_g(x)$, and probability function $p : \mathcal{X} \to [0, 1]$, the functional specified complexity kardis is*

$$\kappa(x) \coloneqq |\mathcal{X}|(1 + \ln|\mathcal{X}|)\frac{p(x)}{F_g(x)^{-1}}.$$

Given the functional specified complexity kardis, the functional specified complexity (FSC) is thus

$$\begin{aligned} FSC(x) &\coloneqq -\log_2\left[|\mathcal{X}|(1 + \ln|\mathcal{X}|)\frac{p(x)}{F_g(x)^{-1}}\right] \\ &= -\log_2 r\frac{p(x)}{\nu(x)} \end{aligned}$$

where we have defined $r = |\mathcal{X}|(1 + \ln|\mathcal{X}|)$ and $\nu(x) = F_g(x)^{-1}$.

Given the above definitions, Montañez [1] upper bounds the probability of observing an outcome with specified complexity at least as extreme as $SC(x)$, as follows.

Theorem 1 (Conservation of Canonical Specified Complexity [1])**.** *Let $p(x)$ be any discrete or continuous probability measure on space $\mathcal{X}$, let $r \in \mathbb{R}_{>0}$ be a scaling constant, and let $\nu : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be any nonnegative integrable function where $\nu(\mathcal{X}) \leq r$. Then*

$$\Pr\left(-\log_2 r\frac{p(X)}{\nu(X)} \geq b\right) \leq 2^{-b},$$

where $X \sim p$.

This probabilistic bound allows one to construct hypothesis tests using the specified complexity value of a given object b. Given a proposed probability measure p, we let $X \sim p$. Theorem 1 then allows for the calculation of 2^{-b} which behaves similar to a p-value, the probability of observing another object at least as extreme as the one in question. Should this value be less than a given significance level α, we conclude that b is a statistically significant specified complexity value and that p is not a plausible explanation for the observation b.

Previous work has used this hypothesis testing framework, called specified-complexity or *two-distribution* hypothesis testing, for applications such as endowing an artificial agent

with intention perception [2] and identifying bias in machine learning training data [3]. These hypothesis tests are described as *two-distribution* hypothesis tests, since they make use of both a probability distribution $p(x)$ and a specification distribution $\nu(x)$. While successful, a severe limitation of these methods is that they require manual computation of specification function values, namely $\nu(x)$. Even in the case of FSC, previous researchers using two-distribution hypothesis tests have had to compute functional specificity, $F_g(x)$, either by brute force or clever combinatorics in order to cover the billions of possible specification values [2], [3]. Moreover, the computation of the specification function was unique to the specific problem at hand, meaning that, regardless of the computational efficiency of the method, each new application of two-distribution hypothesis tests could potentially require a new method for quickly computing the specification function. Our work addresses this current limitation of two-distribution methods by proving finite-sample probabilistic bounds for such hypothesis tests so that practitioners may simply sample to form estimates without knowledge of the specification distribution, eliminating the need to compute exact specification function values.

III. Results

We present our main result first, which upper bounds the probability of observing an object X with specified complexity at least as extreme as $SC(x)$, similar to Theorem 1. However, unlike Theorem 1, ours only requires an estimate of the specification function obtained via sampling, $\hat{\nu}(x)$, rather than the exact value, $\nu(x)$. As such, there is no requirement that the exact specification distribution be known in advance. Lastly, this bound may be used similarly to Theorem 1 to conduct two-distribution hypothesis tests given a significance level α, wherein a null hypothesis explanation for the observed object may be rejected if $\Pr(SC(X) \geq SC(x)) \leq \alpha$.

Theorem 2 (Estimated Specificity Bound)**.** *Let*

$$SC(x) = -\log_2 \kappa(x) = -\log_2 r\frac{p(x)}{\nu(x)}$$

where $p(x)$ is any discrete or continuous probability measure on space $\mathcal{X}$, $r \in \mathbb{R}_{>0}$ is a scaling constant, and $\nu : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is any nonnegative integrable function such that $\nu(\mathcal{X}) \leq r$. Let $\hat{\nu}$ be an estimate for ν where, with probability $1-\delta$, $|\nu(x) - \hat{\nu}(x)| < \epsilon$. Then,

$$\Pr(SC(X) \geq SC(x)) \leq r\frac{p(x)(1-\delta)}{\hat{\nu}(x) - \epsilon} + \delta.$$

Just as FSC provides a concrete way to compute $\nu(x)$, a modified version of Theorem 2 may be attained to provide a more straightforward way to compute the probabilistic bound of Theorem 2.

Corollary 1 (Functional Specificity Bound)**.** *Let*

$$FSC(x) = -\log_2 \kappa(x) = -\log_2 \left[|\mathcal{X}|(1+\ln|\mathcal{X}|)F_g(x)p(x)\right]$$

where $p(x)$ is a probability mass function on discrete finite space $\mathcal{X}$, and $F_g : \mathcal{X} \to (0,1]$ is the functional specificity. Let $\hat{F}_g$ be an estimate for F_g where, with probability $1-\delta$, $|F_g(x) - \hat{F}_g(x)| < \epsilon$. Then,

$$\Pr(FSC(X) \geq FSC(x)) \leq rp(x)(\hat{F}_g(x) + \epsilon)(1-\delta) + \delta$$

where $r = |\mathcal{X}|(1+\ln|\mathcal{X}|)$. When $\hat{F}_g(x)$ is a Bernoulli parameter estimate attained using a sample size of n, this becomes

$$\Pr(FSC(X) \geq FSC(x)) \leq$$
$$|\mathcal{X}|(1+\ln|\mathcal{X}|)p(x)\left(\hat{F}_g(x) + \frac{|Z_{\delta/2}|}{2\sqrt{n}}\right)(1-\delta) + \delta,$$

where $Z_{\delta/2}$ is the standard normal z-score for $\delta/2$.

Notice that Corollary 1 essentially replaces the $\hat{\nu}(x) - \epsilon$ term of Theorem 2 with a quantity only involving a sampled Bernoulli parameter estimate, the number of samples taken, and a z-score. The z-score may also be treated as a parameter of the probabilistic bound, as it is directly related to the δ term. That is, once one is chosen, the other may be calculated directly. We note that as δ decreases, the upper bound decreases as well. However, since $0 \leq \delta \leq 1$, $|Z_{\delta/2}|$ also increases, which can increase the upper bound if $\hat{F}_g(x) + |Z_{\delta/2}|/(2\sqrt{n})$ is large. Thus, changing the value of δ can lead to trade-offs, but these can be offset by increasing the number of samples taken, n.

It is important to note that this probabilistic bound introduces two new parameters compared to Theorem 1. These parameters have a direct impact on the tightness of the bound, and it is essential they are are chosen beforehand in order to avoid "SC-hacking," an equivalent exploitation as p-hacking in traditional hypothesis testing (see Section V).

We also introduce a probabilistic bound on the weighted sum of multiple specified complexity values, which allows one to combine multiple specified complexity models.

Corollary 2 (Combined Specificity Estimation)**.** *Let $SC_1(x), \ldots, SC_m(x)$ be canonical specified complexity models sharing a common probability function p(x). Define a set of mixture variables*

$$\Lambda = \{\lambda_i : i = 1, \ldots, m, 0 \leq \lambda_i \leq 1\},$$

such that $\Sigma_{i=1}^m \lambda_i = 1$. Then for any such Λ,

$$\Pr\left(\sum_{i=1}^m \lambda_i SC_i(X) \geq SC(x)\right) \leq \sum_{i=1}^m \Pr(SC_i(X) \geq SC(x)) \tag{1}$$

where each $P(SC_i(X) \geq SC(x))$ is calculated using Theorem 2.

Lastly, we present a similar bound as Theorem 2 which only requires a sampled estimate of the normalizing constant $\hat{r}$ instead of the estimated specificity term $\hat{\nu}(x)$.

Theorem 3 (Probability Under $\hat{r}$ for Bounded Specification Function)**.** *Let*

$$SC(x) = -\log_2 \kappa(x) = -\log_2 r\frac{p(x)}{\nu(x)}$$

and

$$\hat{r} := |\mathcal{X}| \left(\frac{1}{n} \sum_{i=1}^{n} \nu(x_i) \right)$$

where $x_i, \ldots, x_n$ is a set of n i.i.d. samples drawn uniformly at random from space $\mathcal{X}$. Assume $\nu(X) < k$ almost surely for some positive scalar k. Then for any real number $\epsilon > 0$,

$$\Pr\left(SC(X) \geq SC(x)\right) \leq \left((\hat{r} + \epsilon|\mathcal{X}|) \frac{p(x)}{\nu(x)} \right) [1 - 2e^{-2n\epsilon^2/k^2}] + 2e^{-2n\epsilon^2/k^2}.$$

IV. Examples

We apply the results of the previous section to three different scenarios, demonstrating the effectiveness of the finite-sample bounds and illustrating potential use cases. The first scenario involves a novel method for detecting cheating students on an exam. We demonstrate how each of the components of Corollary 1 may be calculated and thus show how our finite-sample bounds may be used to conduct a two-distribution hypothesis test even when exact specified complexity (SC) bounds have not yet been computed. The remaining two examples recreate scenarios from the existing literature, allowing us to compare our bounds to exactly-computed analytical SC bounds.

The examples all follow a similar pattern. In particular, we begin with an element $x \in \mathcal{X}$, a null hypothesis proposed distribution P such that $X \sim P$, and a significance level α. We then bound $\Pr(SC(X) \geq SC(x))$, the likelihood of finding another element in $\mathcal{X}$ that is more extreme than x under distribution P, using Corollary 1. If this probability is less than α we reject the null hypothesis that P is a probable explanation for x, and otherwise fail to reject this null hypothesis. For the final two examples, we review how the previous exact bounds were computed, compare this process with the simpler sampling framework used for computing the Corollary 1 estimate, and then compare the two bounds. For all examples, we use $\alpha = 0.05$, $\delta = 0.01$ (which corresponds to $|Z_{\delta/2}| = 2.576$), and $n = 10{,}000{,}000$ samples.

A. Detecting Student Cheating

Our first scenario considers an instructor who is interested in determining the likelihood that a group of students cheated on a multiple-choice test through dishonest collaboration. In particular, the instructor is interested in the probability that two students' exams would be as similar as an observed pair by pure chance. Calculating this probability using Corollary 1 first requires us to define the extremity function $g(x)$. In practice, this should capture extremity in an object along an axis that a practitioner wishes to investigate. For our scenario, $g(x)$ should represent the degree of similarity between two students' exams. While tempting to define $g(x)$ based on the number of questions for which a pair of students gave the same answer, several well-performing students may have the same correct answers on many questions. Rather, one would be more suspicious of illicit collaboration between students if their *incorrect* answers matched. Thus, we define $g(x)$ as the ratio of the number of questions for which two students gave the same wrong answer to the total number of questions both students got wrong (regardless of which wrong selection they made). For this hypothetical scenario, we implicitly assume that wrong answers are more or less uncorrelated between honest students.

With this definition of $g(x)$ in mind, the next term we need to compute in order to use Corollary 1 is the estimated proportion of objects more extreme than the observed one, $\hat{F}_g(x)$. In this scenario, this is the estimated proportion of all possible multiple choice test submissions which are more similar to a reference student's exam than the other non-reference student. The choice of which student we use as the reference is arbitrary. In practice, the most straightforward way to compute $\hat{F}_g(x)$ is to uniformly sample n random multiple-choice test submissions. That is, we generate n test submissions whose answers to each question are random. For each of these tests, we compute $g(x)$ and count the number which have a greater $g(x)$ value than the non-reference student. This count divided by the number of samples n yields our desired $\hat{F}_g(x)$. If no randomly generated tests have a greater $g(x)$ value than the non-reference student, we simply set $\hat{F}_g(x) = 1/n$. One could also use Good-Turing frequency estimation in this case [6], but we found no significant difference in results using this method.

The last terms to calculate in order to conduct our two-distribution hypothesis test are $|\mathcal{X}|$ and $p(x)$. The former is simply the number of possible objects in the space of interest. In our case, this may be calculated by raising the number of choices per question to the number of questions. For $p(x)$, we take into account that there is some large probability that a student will answer a question correctly, which for this test, is historically 92% on average. When a student gets a question wrong, they do so with equal probability across the four remaining (wrong) answer choices. Assuming independence among questions, this gives us a weighted "coin-flip" model for student submissions, which is weighted heavily towards correct answers (increasing the number of matching questions expected), but which does not affect our specificity model. We note that even a simplified naïve model that assigns uniform probability across all five answer choices would work for this hypothetical example.

With all of the above terms defined, we may conduct our two-distribution hypothesis test. To illustrate, we create an extreme example in which two students each submit a fifty-question, five-choice test, both get the same twenty questions wrong, and their answers to those twenty questions are all the same. This gives $|\mathcal{X}| = 5^{50}$ and $p(x) = 0.92^{30}(0.08/4)^{20}$. Lastly, our sampling empirically yielded $\hat{F}_g(x) = 1/n = 1/10{,}000{,}000$ and applying Corollary 1 gives

$$\begin{aligned} &P(SC(X) \geq SC(x)) \\ &\quad \leq |\mathcal{X}|(1 + \ln|\mathcal{X}|)p(x) \left(\hat{F}_g(x) + \frac{|Z_{\delta/2}|}{2\sqrt{n}} \right)(1 - \delta) + \delta \\ &\quad = 0.0351. \end{aligned}$$

Since this is less than our significance level of $\alpha = 0.05$, we

reject the null hypothesis that random chance is a plausible explanation for the similarity in two students' exam submissions.

We can also explore a scenario where an instructor is suspicious of a group of three students. Two of these students have the same test submissions as in the previous example. However, the third student also gets twenty questions wrong, but only fifteen of them are the same incorrect answers as the other students. In this case, we may use the combined SC estimation from Corollary 2. Using the same $|\mathcal{X}|$, $p(x)$, and $g(x)$ definitions from above as well as the same sampling method for $\hat{F}_g(x)$ yields $P(SC(X) \geq SC(x)) \leq 0.0702$. Since this bound is greater than our significance level of $\alpha = 0.05$, we fail to reject the null hypothesis that random chance is a plausible explanation for the three students' exam submissions. This aligns with the intuition that an instructor should be less confident that a larger group of students are all illicitly collaborating.

B. Intention Perception in Artificial Agents

Our second scenario involves an artificial agent (depicted as a gopher) whose objective is to survive without being caught by traps. The gopher is equipped with *intention perception*, the ability to detect whether a given configuration of components is likely to be a trap left behind by another agent intending to cause harm, or rather simply a random configuration that looks like a trap. In practice, the intention-perception algorithm is implemented by performing two-distribution hypothesis tests. When presented with a configuration of components, the gopher asks the question, *"What is the probability that I find another configuration that is at least as coherent (i.e., dangerous) as this one?"* If this probability is sufficiently low, the gopher agent will reject the null hypothesis that the observed configuration is a random assortment of components and will conclude that it was created with the intention of harm. However, if the probability is high, the gopher will enter the configuration so that it can eat the food within, extending its life [2].

In this scenario, we use FSC to model the gopher's perceived danger level for a given configuration. We let the extremity function $g(x)$ be the number of coherent connections per non-empty cell in a twelve-cell configuration. A coherent connection is defined as a connection between two wires with the same thickness, at the correct orientation. For instance, the configuration on the left in Figure 1 has four coherent connections, one between each of the projectile-firing end pieces (called arrows) and the wires, and one between each of the wires and the door, while the configuration on the right has one coherent connection, between the wires at the top left of the configuration. A non-empty cell is defined as any cell that has a wire or arrow. The configuration on the left in Figure 1 has four nonempty cells, while the figure on the right has nine. Thus, for the configuration on the left, $g(x) = 4/4 = 1$, and for the configuration on the right, $g(x) = 1/9$.

Previous work [2] has used Theorem 1 to directly compute $\Pr(SC(X) \geq SC(x))$. As discussed in Section II, this requires the calculation of $F_g(x) = M_g(x)/|\mathcal{X}|$, the proportion of possible configurations more extreme than the observed one. To compute $F_g(x)$, Hom et al. [2] pre-computed $M_g(x)$, the number of configurations with a level of function greater than or equal to that of x, for every possible ratio $g(x)$ of coherent connections to nonempty cells. Using these methods, the bound obtained using the exact specificity of each configuration is $\Pr(SC(X) \geq SC(x)) \leq 2.6 \times 10^{-12}$ for the configuration on the left and $\Pr(SC(X) \geq SC(x)) \leq 34.4$ for the configuration on the right. Note that although the result is greater than 1 for the figure on the right, this simply reflects the lack of structure in the configuration. Therefore, when $\alpha = 0.05$, the gopher agent would reject the null hypothesis that the configuration was randomly generated for the trap on the left, but fail to reject it for the configuration on the right, allowing the gopher to selectively enter configurations.

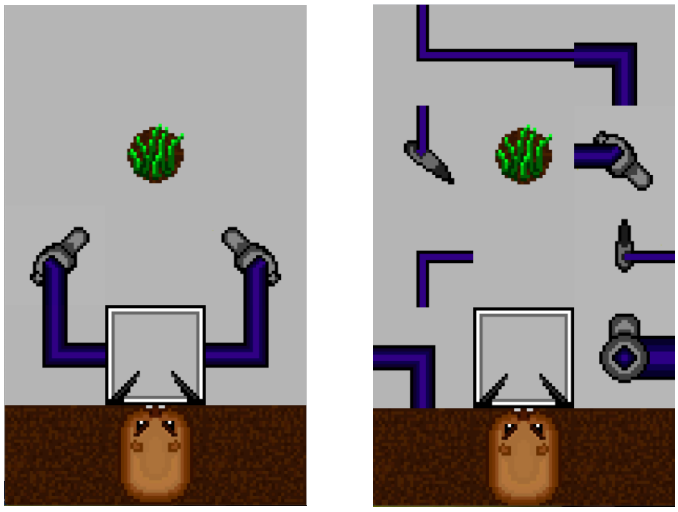

Fig. 1: Two configurations in the simulated gopher world.

While computing the exact value of $M_g(x)$ allows us to use Theorem 1 to perform two-distribution hypothesis tests and ultimately determine whether a given trap configuration is likely to be dangerous, this method may not always be feasible. In this scenario, we assumed that the gopher agent had a knowledge of the entire space of possible configurations. However in practice, an artificial agent will likely have only seen some subset of these configurations. Additionally, even if the agent does have a knowledge of the entire space of configurations, configurations of larger dimensions than the one in this scenario would likely require much more computation.

Therefore, we now show how one can still use two-distribution hypothesis tests to model the intention perception algorithm with unknown $\nu(x)$. Instead of pre-computing $M_g(x)$, we instead generate a uniform random sample of configurations and count the number of these configurations that are more extreme than x to obtain the estimate $\hat{\nu}(x)$. After using the same methods as [2] to compute $|\mathcal{X}|$ and $p(x)$, we can then apply Corollary 1 to determine the likelihood that

this configuration is dangerous to the agent.

We now apply this process to the configuration on the left of Figure 1. This configuration has four coherent connections and four nonempty cells, so $g(x) = 4/4 = 1$. Thus, applying Corollary 1 gives $\Pr(SC(X) \geq SC(x)) \leq 0.0268$. Since $0.0268 < \alpha$, we reject the null hypothesis that the configuration was randomly generated in favor of the alternative hypothesis that the configuration was designed by an agent who intended to cause harm to the gopher.

We can also apply the same process to a randomly generated configuration, such as the one on the right in Figure 1. This configuration has one coherent connection and nine nonempty cells, so $g(x) = 1/9$. Applying the same process as before, we conduct our two-distribution hypothesis test to obtain $\Pr(SC(X) \geq SC(x)) \leq 32.8$. Since $32.8 > \alpha$, we fail to reject the null hypothesis that the trap was randomly generated. Note that Corollary 1 gives an upper bound on $\Pr(SC(X) \geq SC(x))$ rather than the true probability itself. Thus, the fact that the bound is greater than 1 simply reflects that the configuration does not exhibit large degrees of coherence.

Despite only using the estimated value $\hat{\nu}(x)$, the results are strikingly consistent with those of Hom et al. [2]. These examples demonstrate that two-distribution hypothesis tests can still equip an agent with intention perception even when the agent does not have the knowledge to compute the exact value of $\nu(x)$.

C. Identifying Bias in Data

While the gopher agent in the previous section uses two-distribution hypothesis tests to distinguish random and designed traps, these hypothesis tests may also be used to detect representational bias in tabular machine learning datasets [3]. That is, given some proposed fair distribution, two-distribution hypothesis tests may be used to measure the deviation of another dataset from it, a measure of potential bias. Importantly, this may be done without the need to first train a model on the potentially-biased data to analyze its output.

As in Yik et al. [3], we test the well-known COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset [7] for bias using two-distribution hypothesis tests. This dataset is used to train machine learning models to predict the recidivism risk (*low, medium, high*) for a given offender based on certain features such as their race, gender, and type of offense. It has been previously shown that models trained on this dataset without specific corrective measures tend to label a disproportionate amount of African-American people as having a high risk of recidivism [7]–[10].

Previous work [3] defined the extremity function $g(x)$ as the ℓ_1 (city block) metric and used this function to capture the distance from one distribution of recidivism risks to another. In this instance, we are interested in the distance $g(x)$ between the recidivism risk distribution of African-Americans from that of Caucasians and the number of possible datasets which could produce a more extreme $g(x)$. That is, the number of datasets (of the same shape as COMPAS) which would be considered at least as unfair as COMPAS. With this $g(x)$ function, Yik et al. [3] used a specialized combinatorial formula to systematically count the number of possible datasets with a greater $g(x)$ value than COMPAS, namely $M_g(x)$.

While useful, such a specialized routine for computing $M_g(x)$ and subsequently $\nu(x)$ may not be applicable for other use cases of two-distribution hypothesis tests, and practitioners may not have the expertise to derive such combinatorial counting schemes. As such, we demonstrate that a two-distribution test may still be conducted by only using the sampled $\hat{\nu}(x)$. Instead of computing $M_g(x)$ by counting the number of possible datasets with a greater $g(x)$ value than COMPAS, we adopt a simple sampling process as in the previous subsections. Specifically, we uniformly at random generate possible recidivism risk score distributions for African-Americans, thereby estimating the number of possible datasets of the same size as COMPAS which would be considered more unfair. Using this, we calculate $\hat{\nu}(x)$, our estimate for the specification, as in the previous subsection. Finally, using the same methods for computing $|\mathcal{X}|$ and $p(x)$ as [3], we conduct our two-distribution hypothesis test using Corollary 1 to obtain $\Pr(SC(X) \geq SC(x)) \leq 0.0100$. While relatively large compared to the exact probability bound of 2.4×10^{-44} reported in Section 7.2 of Yik et al. [3], we find that the bound is lower than our significance level of $\alpha = 0.05$, so we reject the null hypothesis regardless. Thus, while only sampling a small fraction of all possible datasets in $\mathcal{X}$, we are able replicate the hypothesis test result of Yik et al. [3], showing that two-distribution hypothesis tests remain viable even when the exact distribution of $\nu(x)$ is not known.

V. Discussion

The examples from Section IV demonstrate the utility of our finite-sample bounds in detecting student cheating, providing intention-perception to artificial agents, and identifying bias in data. The bounds computed in all three examples are consistent with the results of prior work, showing that finite-sample bounds allow one to use the two-distribution hypothesis tests introduced by Montañez [1] without incurring heavy computational costs or requiring complex combinatorics. Furthermore, while empirical p-values require sampling from the true distribution under consideration, our two-distribution hypothesis tests do not. This gives a major advantage in situations where sampling from the true distribution is costly or impossible.

In particular, applications such as identifying genetically modified DNA, distinguishing the outputs of artificial intelligence from those of humans, and detecting malicious network activity all involve large sample spaces or a significant possibility of errors in sampled data, making it difficult or impossible to compute exact quantities. When examining genetically modified DNA, for instance, researchers likely only have a small subset of the DNA sequences of a given species. The role of physical equipment in DNA sequencing also introduces the potential for errors, such that it may be difficult for researchers to be certain that they are sampling from the true distribution. When distinguishing the output of

artificial general intelligence from that of humans, we likely won't be able to compute metrics on all of human or artificially generated text or video. Similarly, when detecting malicious network activity, one can only access some sample of global network activity. Therefore, our finite-sample bounds broaden the range of applications for two-distribution hypothesis tests.

One might note that although the bounds given in Section IV match the results of prior work, the bounds obtained are higher than the exact bounds given by existing methods [1]. However, this simply reflects the greater amount of uncertainty of using a finite sample as opposed to computing exact quantities. To reduce this uncertainty, one can take a larger sample size or change their choice of the δ parameter.

A. Parameter Choice for δ and n

When using Corollary 1 to conduct two-distribution hypothesis tests, the user may choose any values for δ and n such that $0 \leq \delta \leq 1$ and $n \in \mathbb{Z}^+$. For the experiments in Section IV, we chose $\delta = 0.01$ and $n = 10{,}000{,}000$. However, it may make sense to choose a different values for δ and n depending on the particular application.

The effect of δ and n on the error of the bound is shown in Figure 2. To calculate the error, we compare the Corollary 1 estimate with the exact bound given by Theorem 1. For fixed δ, the error of the bound decreases as n, the number of samples, increases. This is expected, as using a larger proportion of the sample space should cause $|Z_{\delta/2}|/(2\sqrt{n})$ to decrease for a given confidence level $1 - \delta$. Thus, in order to obtain a tight bound, it is advantageous to sample a large number of instances n. Ultimately, the value of n that one chooses will be a trade-off between decreasing $|Z_{\delta/2}|/(2\sqrt{n})$ and the cost of obtaining more samples.

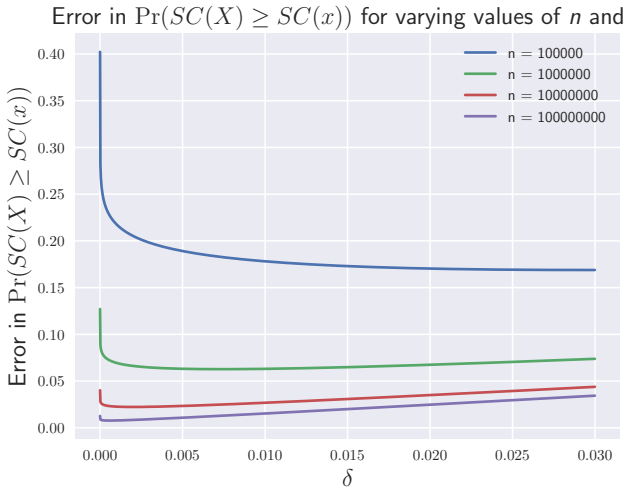

Fig. 2: The error in $\Pr(SC(X) \geq SC(x))$ for a trap with $g(x) = 4/4$ varies with respect to n and δ.

As δ increases for any fixed n, the curve representing the error of the bound first decreases and then increases. To understand why this is the case, consider varying δ and holding all other quantities constant. Note that $Z_{\delta/2}$ is not constant, as $Z_{\delta/2}$ is the z-score corresponding to $\delta/2$. As stated in Corollary 1, $1 - \delta$ is the probability that $|F_g(x) - \hat{F}_g(x)| < \epsilon$. The bound given by Corollary 1 therefore consists of two terms, the first which encompasses the contribution to the probability for the case where $|F_g(x) - \hat{F}_g(x)| < \epsilon$, and the second which encompasses this contribution for the case where $|F_g(x) - \hat{F}_g(x)| \geq \epsilon$. As δ increases, both $1 - \delta$ and $|Z_{\delta/2}|/(2\sqrt{n})$ decrease, so the first term of the bound decreases while the second increases. As δ decreases, the opposite occurs. The trade-off between minimizing these two terms of the equation is reflected by the shape of the curves in Figure 2. In particular, the minimum of each curve is the "optimal" value for δ that gives the bound with the smallest error.

Given that there is an "optimal" value for δ, one may be tempted to take a sample of n elements, observe the estimated value $\hat{\nu}(x)$, and then try a range of δ values to find the tightest possible bound. Or, after failing to reject the null hypothesis, one might be tempted to increase n to lower the bound. However, both of these are forms of "SC-hacking", the practice of adjusting the parameters of a specified complexity hypothesis test after already having observed the data in order to report a more significant result [1]. SC-hacking is analogous to p-hacking and allows for the influence of bias as well as false positives, or the incorrect detection of anomalies where there are none.

To avoid SC-hacking, one can instead use a "placeholder" value for $\hat{\nu}(x)$ to try various values of δ and n *before* observing the true value of $\hat{\nu}(x)$. This is a similar to estimating the effect size, the difference in group means, in an intervention-based study to determine the sample size n required to avoid Type II errors [11]. Although effect sizes in the traditional sense do not apply directly to our work, as our hypothesis tests do not involve control and intervention groups, using a placeholder value for $\hat{\nu}(x)$ plays a similar role to estimating an effect size by allowing us to choose suitable values for δ and n and reduce Type II errors.

After finding a placeholder value for $\hat{\nu}(x)$, one might suggest finding the first derivative of the resulting bound with respect to δ in order to choose the δ parameter that gives the tightest bound. However, note that $Z_{\delta/2}$ is the z-score that corresponds to a p-value of $\delta/2$, that is, $Z_{\delta/2} = \Phi^{-1}(\delta/2)$. We cannot differentiate $\Phi^{-1}(\delta/2)$ with respect to δ since Φ^{-1}, the probit function, does not have a closed-form expression [12]. Therefore, in practice, we suggest either using numerical methods such as gradient descent or simply testing a range of values for δ and choosing the one that gives the tightest bound.

Beyond malicious tampering of the δ and n parameters, SC hypothesis tests, like their traditional p-value contemporaries, are also prone to hacking via multiple hypothesis testing. Although our work only tests single hypotheses, a malicious practitioner could, for example, preprocess the data of Section IV-C and conduct an SC test multiple times until a significant result is seen. In order to avoid such SC-hacking, users of our SC tests seeking to explore multiple hypotheses must proceed with the same caution they would have when testing multiple

hypothesis with a traditional test. Correction methods for multiple hypothesis testing include the Bonferroni correction [13], Tukey's honestly significant difference (HSD) test [14], and the Benjamini-Hochberg procedure [15].

VI. Related Work

Anomaly detection is the practice of identifying data points that deviate from the normal behavior of the rest of the dataset. Chandola et al. give a survey of anomaly detection techniques, including classification, clustering, nearest neighbor and density methods, statistical methods, information theoretic techniques, and deep learning [16]–[20]. These methods have been applied widely across various domains, including computer networking [21], [22], DNA sequencing [23], [24], and detecting artificially generated text [25], [26].

Our work builds on previous work done in statistical anomaly detection. In particular, Montañez reduced specified complexity testing to a form of statistical hypothesis testing, demonstrating that one can conduct a two-distribution hypothesis test by comparing the canonical specified complexity kardis to an α value [1]. This form of hypothesis testing allows one to consider the specificity, or degree of structure, of an instance, in addition to the complexity, or how likely the instance is to be selected from a given space. Importantly, these two-distribution hypothesis tests are distinct from p-values, as they can be used for arbitrary probability distributions. Specified complexity as a mathematical concept was introduced by Dembski [4], and subsequently refined by Dembski, Marks, Ewert, and others [27]–[31].

The hypothesis tests proposed by Montañez [1] have been shown to be useful in multiple contexts. Díaz-Pachón, Hössjer, and their collaborators have proposed two-distribution hypothesis tests for a variety of applications, ranging from epistemology to biology [32]–[35]. Hom et al. used these hypothesis tests to provide artificial gopher agents with a form of intention perception, allowing the gopher agents to distinguish randomly generated traps from ones that were intentionally designed to hurt the gopher [2]. Yik et al. applied the tests to the well-known COMPAS dataset, and found that the dataset has bias in African-American recidivism scores [3]. However, in prior studies, the exact specificity value for each instance had to be calculated using complex combinatorics and numerical methods in order to conduct the two-distribution hypothesis test. Our work provides bounds that allow one to use the hypothesis tests originally proposed by Montañez [1] without the need to compute exact specificity values, reducing the need for these computations and significantly broadening the range of use cases for these hypothesis tests.

VII. Conclusion

The ability to distinguish normal from extreme, fair from unfair, and dangerous from safe has become increasingly important in today's data-driven world. Two-distribution hypothesis tests, which capture both the probability of observing and object and its degree of structure, have proved to be a useful tool for detecting such anomalies and have previously been applied to endow artificial agents with intention perception [2] and identify bias in machine learning training data [3]. These hypothesis tests rely on calculating the probability that a *specified complexity* value more extreme than that of a given object would be observed at random. However, calculating the specificity portion, $\nu(x)$, required by these tests can often be tedious, time-consuming, and specific to the task at hand, due to the requirement of brute-force computation for each $\nu(x)$ value or the need to design a custom combinatorial method to avoid this computation. As such, practitioners hoping to use two-distribution hypothesis tests may be limited by the difficulty of analytically calculating the specificity term. We present a set of finite-sample probabilistic bounds for two-distribution hypothesis tests which only require a sampled estimate, $\hat{\nu}(x)$, of the specificity value. This lowers the barrier for using two-distribution tests from analytically computing the specificity term to deriving a simple sampling procedure for objects in the problem space. A similar bound for two-distribution tests which only requires a sampled estimate of the normalizing constant r is also presented. We apply our new bounded hypothesis tests to three scenarios. In the first, we explore how an instructor may detect unpermitted student collaboration on a test, illustrating how practitioners may make use of the new hypothesis tests in their work. In the remaining two applications, we recreate scenarios from the literature [2], [3] and compare our new finite-sample bounded results to the exactly-computed values. Overall, we find that two-distribution hypothesis tests may still be conducted effectively using our probabilistic bounds in place of analytical values. Lastly, we explore the effect of the δ parameter and sample size n on our bounded hypothesis tests and again warn against the possibility of statistical misuse.

Future work includes exploring scenarios where the complexity term $p(x)$ may be difficult to compute analytically instead of the specificity term $\nu(x)$ or normalizing constant r. Similarly bounded hypothesis tests that only require a sampled estimate of $p(x)$ may also prove useful for practitioners wishing to apply two-distribution hypothesis tests to novel scenarios such as distinguishing text generated by a large language model from work written by humans and detecting malicious network activity. Furthermore, we primarily explore functional specified complexity in this work, but other types of specified complexity such as quantitative irreducible complexity [1], [36] may also be worth considering. Exploring possible applications of our various bounded two-distribution hypothesis tests within these frameworks may open new avenues for identifying anomalous artifacts.

VIII. Acknowledgements

This research was supported in part by the National Science Foundation under Grant No. 1950885 and Global Scholars. Any opinions, findings, or conclusions are those of the authors alone, and do not necessarily reflect the views of the National Science Foundation or Global Scholars.

References

[1] G. D. Montañez, "A Unified Model of Complex Specified Information," *BIO-Complexity*, vol. 2018, 2018.

[2] C. Hom, A. Maina-Kilaas, K. Ginta, C. Lay, and G. Montañez, "The Gopher's Gambit: Survival Advantages of Artifact-based Intention Perception," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, vol. 1, 2021.

[3] W. Yik, L. Serafini, T. Lindsey, and G. D. Montañez, "Identifying Bias in Data Using Two-Distribution Hypothesis Tests," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 831–844.

[4] W. A. Dembski, *The Design Inference: Eliminating Chance through Small Probabilities*. Cambridge: Cambridge University Press, 1998.

[5] R. M. Hazen, P. L. Griffin, J. M. Carothers, and J. W. Szostak, "Functional Information and the Emergence of Biocomplexity," *Proceedings of the National Academy of Sciences*, vol. 104, no. suppl_1, pp. 8574–8581, 2007.

[6] W. A. Gale and G. Sampson, "Good-Turing Frequency Estimation Without Tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," in *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264.

[8] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[9] C. DiCiccio, S. Vasudevan, K. Basu, K. Kenthapadi, and D. Agarwal, "Evaluating Fairness using Permutation Tests," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1467–1477.

[10] B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen, "A Statistical Test for Probabilistic Fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 648–665.

[11] G. M. Sullivan and R. Feinn, "Using effect size—or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.

[12] K. E. Train, *Discrete choice methods with simulation*. Cambridge university press, 2009.

[13] P. Sedgwick, "Multiple significance tests: the bonferroni correction," *Bmj*, vol. 344, 2012.

[14] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.

[15] D. Thissen, L. Steinberg, and D. Kuang, "Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons," *Journal of educational and behavioral statistics*, vol. 27, no. 1, pp. 77–83, 2002.

[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[17] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially Learned Anomaly Detection," in *2018 IEEE International conference on data mining (ICDM)*. IEEE, 2018, pp. 727–736.

[18] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[19] A. Patcha and J.-M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends," *Computer networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[20] T. Zhan and J. Kang, "Finite-sample two-group composite hypothesis testing via machine learning," *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, vol. 31, no. 3, pp. 856–865, 2022.

[21] S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri, "Self-Nonself Discrimination in a Computer," in *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*. IEEE, 1994, pp. 202–212.

[22] M. Ahmed, A. Naser Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804515002891

[23] A. Milosavljević, "Discovering Dependencies via Algorithmic Mutual Information: A Case Study in DNA Sequence Comparisons," *Machine Learning*, vol. 21, pp. 35–50, 1995.

[24] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection for Discrete Sequences: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 823–839, 2012.

[25] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," 2019.

[26] M. M. Dalkilic, W. T. Clark, J. C. Costello, and P. Radivojac, "Using Compression to Identify Classes of Inauthentic Texts," in *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 2006, pp. 604–608.

[27] W. A. Dembski, "Specification: The Pattern that Signifies Intelligence," *Philosophia Christi*, vol. 7, no. 2, pp. 299–343, 2005.

[28] W. Ewert, W. A. Dembski, and R. J. Marks II, "Algorithmic Specified Complexity," *Engineering and Metaphysics*, 2012.

[29] W. Ewert, R. J. Marks, and W. A. Dembski, "On the Improbability of Algorithmic Specified Complexity," in *System Theory (SSST), 2013 45th Southeastern Symposium on*. IEEE, 2013, pp. 68–70.

[30] D. Nemati and E. Holloway, "Expected Algorithmic Specified Complexity," *BIO-Complexity*, vol. 2019, 2019.

[31] J. Bartlett, "Active Information is a Specified Complexity Model," *Communications of the Blyth Institute*, vol. 2, no. 2, pp. 40–41, 2020.

[32] D. A. Díaz-Pachón, J. P. Sáenz, and J. S. Rao, "Hypothesis Testing with Active Information," *Statistics & Probability Letters*, vol. 161, p. 108742, 2020.

[33] O. Hössjer, D. A. Díaz-Pachón, and J. S. Rao, "A Formal Framework for Knowledge Acquisition: Going Beyond Machine Learning," *Entropy*, vol. 24, no. 10, p. 1469, 2022.

[34] D. A. Díaz-Pachón and O. Hössjer, "Assessing, Testing and Estimating the Amount of Fine-tuning by Means of Active Information," *Entropy*, vol. 24, no. 10, p. 1323, 2022.

[35] S. Thorvaldsen and O. Hössjer, "Using Statistical Methods to Model the Fine-tuning of Molecular Machines and Systems," *Journal of Theoretical Biology*, vol. 501, p. 110352, 2020.

[36] M. J. Behe, *Darwin's Black Box: The Biochemical Challenge to Evolution*. Simon and Schuster, 1996.

APPENDIX

A. Proofs

Theorem 2 (Estimated Specificity Bound)**.** *Let*

$$SC(x) = -\log_2 \kappa(x) = -\log_2 r\frac{p(x)}{\nu(x)}$$

where $p(x)$ is any discrete or continuous probability measure on space $\mathcal{X}$, $r \in \mathbb{R}_{>0}$ is a scaling constant, and $\nu : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is any nonnegative integrable function such that $\nu(\mathcal{X}) \leq r$. Let $\hat{\nu}$ be an estimate for ν where, with probability $1-\delta$, $|\nu(x) - \hat{\nu}(x)| < \epsilon$. Then,

$$\Pr(SC(X) \geq SC(x)) \leq r\frac{p(x)(1-\delta)}{\hat{\nu}(x) - \epsilon} + \delta.$$

Proof. Let $b := SC(x)$, A denote the event $|\nu(x) - \hat{\nu}(x)| < \epsilon$, and $\overline{A}$ denote the set complement of A. By the Law of Total Probability,

$$\begin{aligned}\Pr(SC(X) \geq b) &= \Pr(SC(X) \geq b \mid A)\Pr(A) \\ &\quad + \Pr(SC(X) \geq b \mid \overline{A})\Pr(\overline{A}) \\ &\leq \Pr(SC(X) \geq b \mid A)\Pr(A) + \Pr(\overline{A})\end{aligned}$$

where the inequality holds because probabilities are upper bounded by 1. By our conditions, $\Pr(A) = 1 - \delta$, giving

$$\Pr\left(SC(X) \geq b\right) \leq \Pr(SC(X) \geq b \mid A)(1-\delta) + \delta. \quad (2)$$

Note that whenever A holds,

$$b = -\log_2 r\frac{p(x)}{\nu(x)} \geq -\log_2 r\frac{p(x)}{\hat{\nu}(x) - \epsilon},$$

thus implying

$$\begin{aligned}\Pr\left(SC(X) \geq b \middle| A\right) &\leq \Pr\left(SC(X) \geq -\log_2 r\frac{p(x)}{\hat{\nu}(x) - \epsilon} \middle| A\right) \\ &\leq r\frac{p(x)}{\hat{\nu}(x) - \epsilon},\end{aligned}$$

where the final inequality follows from application of Theorem 2 of Montañez [1]. Substituting this upper bound into (2), we obtain the desired result. □

Corollary 1 (Functional Specificity Bound)**.** *Let*

$$FSC(x) = -\log_2 \kappa(x) = -\log_2\left[|\mathcal{X}|(1 + \ln|\mathcal{X}|)F_g(x)p(x)\right]$$

where $p(x)$ is a probability mass function on discrete finite space $\mathcal{X}$, and $F_g : \mathcal{X} \to (0, 1]$ is the functional specificity. Let $\hat{F}_g$ be an estimate for F_g where, with probability $1-\delta$, $|F_g(x) - \hat{F}_g(x)| < \epsilon$. Then,

$$\Pr(FSC(X) \geq FSC(x)) \leq rp(x)(\hat{F}_g(x) + \epsilon)(1-\delta) + \delta$$

where $r = |\mathcal{X}|(1 + \ln|\mathcal{X}|)$. When $\hat{F}_g(x)$ is a Bernoulli parameter estimate attained using a sample size of n, this becomes

$$\begin{aligned}&\Pr(FSC(X) \geq FSC(x)) \leq \\ &\quad |\mathcal{X}|(1 + \ln|\mathcal{X}|)p(x)\left(\hat{F}_g(x) + \frac{|Z_{\delta/2}|}{2\sqrt{n}}\right)(1-\delta) + \delta,\end{aligned}$$

where $Z_{\delta/2}$ is the standard normal z-score for $\delta/2$.

Proof. Let $b := FSC(x)$, A denote the event $|F_g(x) - \hat{F}_g(x)| < \epsilon$, and $\overline{A}$ denote the set complement of A. By the Law of Total Probability,

$$\begin{aligned}\Pr(FSC(X) \geq b) &= \Pr(FSC(X) \geq b \mid A)\Pr(A) \\ &\quad + \Pr(SC(X) \geq b \mid \overline{A})\Pr(\overline{A}) \\ &\leq \Pr(FSC(X) \geq b \mid A)\Pr(A) + \Pr(\overline{A})\end{aligned}$$

where the inequality holds because probabilities are upper bounded by 1. By our conditions, $\Pr(A) = 1 - \delta$, giving

$$\Pr\left(FSC(X) \geq b\right) \leq \Pr(FSC(X) \geq b \mid A)(1-\delta) + \delta. \quad (3)$$

Letting $r := |\mathcal{X}|(1 + \ln|\mathcal{X}|)$, we note that whenever A holds,

$$b = -\log_2\left[rp(x)F_g(x)\right] \geq -\log_2[rp(x)(\hat{F}_g(x) + \epsilon)],$$

thus implying

$$\begin{aligned}&\Pr\left(FSC(X) \geq b \mid A\right) \\ &\quad \leq \Pr\left(FSC(X) \geq -\log_2[rp(x)(\hat{F}_g(x) + \epsilon)] \mid A\right) \\ &\quad \leq rp(x)(\hat{F}_g(x) + \epsilon),\end{aligned}$$

where the last inequality follows from application of Theorem 2 of Montañez [1]. Plugging this into (3), we obtain the first desired result.

For the second form, under uniform sampling to estimate the population Bernoulli parameter F_g, we have that with probability $1-\delta$,

$$\begin{aligned}\epsilon &\leq |Z_{\frac{\delta}{2}}|\sqrt{\frac{F_g(x)(1 - F_g(x))}{n}} \\ &\leq |Z_{\frac{\delta}{2}}|\sqrt{\frac{0.5(0.5)}{n}} \\ &= \frac{|Z_{\frac{\delta}{2}}|}{2\sqrt{n}}.\end{aligned}$$

Replacing ϵ with this upper bound gives the second result. □

Remark: Given that $r = |\mathcal{X}|(1+\ln|\mathcal{X}|)$, under a uniform $p(x)$ this bound will require a number of samples on the order of $(\ln|\mathcal{X}|)^2$ before it becomes nontrivial.

Corollary 2 (Combined Specificity Estimation)**.** *Let $SC_1(x), \ldots, SC_m(x)$ be canonical specified complexity models sharing a common probability function p(x). Define a set of mixture variables*

$$\Lambda = \{\lambda_i : i = 1, \ldots, m, 0 \leq \lambda_i \leq 1\},$$

such that $\Sigma_{i=1}^m \lambda_i = 1$. Then for any such Λ,

$$\Pr\left(\sum_{i=1}^m \lambda_i SC_i(X) \geq SC(x)\right) \leq \sum_{i=1}^m \Pr(SC_i(X) \geq SC(x)) \quad (1)$$

where each $P(SC_i(X) \geq SC(x))$ is calculated using Theorem 2.

Proof. Following the proof of Theorem 5 in [1],

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{m}\lambda_i SC_i(X)\geq SC(x)\right)\\
&\qquad\leq \Pr\left(\sum_{i=1}^{m}\lambda_i \max_{i=1,\ldots,m} SC_i(X)\geq SC(x)\right)\\
&\qquad\leq \Pr\left(\max_{i=1,\ldots,m} SC_i(X)\sum_{i=1}^{m}\lambda_i\geq SC(x)\right)\\
&\qquad\leq \Pr\left(\max_{i=1,\ldots,m} SC_i(X)\geq SC(x)\right)\\
&\qquad\leq \Pr\left(\bigvee_{i=1,\ldots,m} SC_i(X)\geq SC(x)\right)\\
&\qquad\leq \sum_{i=1}^{m}\Pr(SC_i(X)\geq SC(x)).
\end{aligned}$$

□

Theorem 3 (Probability Under $\hat{r}$ for Bounded Specification Function)**.** *Let*

$$SC(x) = -\log_2\kappa(x) = -\log_2 r\frac{p(x)}{\nu(x)}$$

and

$$\hat{r} := |\mathcal{X}|\left(\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)\right)$$

where $x_i,\ldots,x_n$ is a set of n i.i.d. samples drawn uniformly at random from space $\mathcal{X}$. Assume $\nu(X)<k$ almost surely for some positive scalar k. Then for any real number $\epsilon>0$,

$$\begin{aligned}
\Pr\left(SC(X)\geq SC(x)\right)\leq &\left((\hat{r}+\epsilon|\mathcal{X}|)\frac{p(x)}{\nu(x)}\right)[1-2e^{-2n\epsilon^2/k^2}]\\
&+2e^{-2n\epsilon^2/k^2}.
\end{aligned}$$

Proof. For any canonical specified complexity model with

$$r=\sum_{x\in\mathcal{X}}\nu(x)$$

we have

$$\begin{aligned}
|\hat{r}-r| &= \left||\mathcal{X}|\left(\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)\right)-|\mathcal{X}|\left(\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\nu(x)\right)\right|\\
&= \left||\mathcal{X}|\left(\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)-\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\nu(x)\right)\right|\\
&= |\mathcal{X}|\left|\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)-\mathbb{E}_{\mathcal{U}_X}[\nu(X)]\right|.
\end{aligned}$$

Let A denote the event that $|\hat{r}-r|\leq\epsilon|\mathcal{X}|$ (where we scale by $|\mathcal{X}|$ since we expect the error to accumulate as the size of the space increases). Invoking Hoeffding's Inequality we obtain

$$\begin{aligned}
1-\Pr(A) &= 1-\Pr\left(|\hat{r}-r|\leq\epsilon|\mathcal{X}|\right)\\
&= \Pr\left(|\hat{r}-r|>\epsilon|\mathcal{X}|\right)\\
&= \Pr\left(|\mathcal{X}|\left|\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)-\mathbb{E}_{\mathcal{U}_X}[\nu(X)]\right|>\epsilon|\mathcal{X}|\right)\\
&= \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\nu(x_i)-\mathbb{E}_{\mathcal{U}_X}[\nu(X)]\right|>\epsilon\right)\\
&\leq 2e^{-2n\epsilon^2/k^2}.
\end{aligned}$$

Furthermore, letting $b:=SC(x)$, we have

$$\begin{aligned}
\Pr\left(SC(X)\geq b\;\middle|\;A\right) &= \Pr\left(\kappa(X)\leq r\frac{p(x)}{\nu(x)}\;\middle|\;A\right)\\
&\leq \Pr\left(\kappa(X)\leq(\hat{r}+\epsilon|\mathcal{X}|)\frac{p(x)}{\nu(x)}\;\middle|\;A\right)\\
&\leq (\hat{r}+\epsilon|\mathcal{X}|)\frac{p(x)}{\nu(x)}
\end{aligned}$$

where the final inequality follows from Corollary 1 of Montañez [1]. Thus, we conclude that

$$\begin{aligned}
\Pr\left(SC(X)\geq b\right) &\leq \Pr\left(SC(X)\geq b\;\middle|\;A\right)\Pr(A)+1-\Pr(A)\\
&\leq \left((\hat{r}+\epsilon|\mathcal{X}|)\frac{p(x)}{\nu(x)}\right)\Pr(A)+1-\Pr(A)\\
&\leq \left((\hat{r}+\epsilon|\mathcal{X}|)\frac{p(x)}{\nu(x)}\right)[1-2e^{-2n\epsilon^2/k^2}]\\
&\quad+2e^{-2n\epsilon^2/k^2}.
\end{aligned}$$

□