Trading query complexity for sample-based testing and multi-testing scalability

Eldar Fischer
Faculty of Computer Science
Israel Institute of Technology
Haifa, Israel
eldar@cs.technion.ac.il

Oded Lachish
Birkbeck, University of London
London, UK
oded@dcs.bbk.ac.uk

Yadu Vasudev
Faculty of Computer Science
Israel Institute of Technology
Haifa, Israel
yaduvasudev@gmail.com

Abstract

We show that every non-adaptive property testing algorithm making a constant number of queries, over a fixed alphabet, can be converted to a sample-based (as per [Goldreich and Ron, 2015]) testing algorithm whose average number of queries is a fixed, smaller than 1, power of n. Since the query distribution of the sample-based algorithm is not dependent at all on the property, or the original algorithm, this has many implications in scenarios where there are many properties that need to be tested for concurrently, such as testing (relatively large) unions of properties, or converting a Merlin-Arthur Proximity proof (as per [Gur and Rothblum, 2013]) to a proper testing algorithm.

The proof method involves preparing the original testing algorithm for a combinatorial analysis. For the analysis we develop a structural lemma for hypergraphs that may be of independent interest. When analyzing a hypergraph that was extracted from a 2-sided test, it allows for finding generalized sunflowers that provide for a large-deviation type analysis. For 1-sided tests the bounds can be improved further by applying Janson's inequality directly over our structures.

Keywords

property testing; sampling; hypergraphs

I. Introduction

A *test* for a property $L \subseteq \Xi^n$ (where Ξ is a fixed finite alphabet), with proximity parameter ϵ, is an algorithm that queries an input $w \in \Xi^n$ in a limited number of places, and distinguishes with high probability between the case where $w \in L$ and the case where no $w' \in L$ is ϵ-close to w in the normalized Hamming distance. A *non-adaptive* test is a test that decides its queries in advance of receiving the corresponding input values, which basically means that its queries are governed by a single distribution μ over the power set of $[n]$.

Given a family of properties $\mathcal{F} \subseteq \{L : L \subseteq \Xi^n\}$, we say that there is a *canonical* testing scheme for $\mathcal{F}$ if there are non-adaptive tests (with the same parameter ϵ) for all members $L \in \mathcal{F}$, which additionally all share the same query probability distribution μ.

This concept has been defined and used before. The most well-known example is that of [1], where the family of all properties of dense graphs (as per the model defined in [2]) with n vertices that are testable (non-adaptively or not) with up to q queries, is shown to have a canonical testing scheme, where the common query distribution consists of uniformly picking a set of $2q$ vertices and querying all $\binom{2q}{2}$ vertex pairs.

Note that being in the dense graph model in essence restricts the admissible properties. Under this model, an input $w \in \{0,1\}^{\binom{n}{2}}$ is interpreted as the adjacency matrix of a graph with n vertices, and a property L is admissible if it is invariant under all input transformations corresponding to re-labeling the graph vertices (i.e., the transformations corresponding to graph isomorphisms).

There are other examples. For example, given a finite field F, for properties of functions over a linear space over F that are known to be invariant under linear transformations, a canonical testing scheme would consist of querying the function over an entire small dimensional subspace picked uniformly at random [3].

A natural question is what would be a candidate for an "ultimate" canonical scheme, where there are no structural impositions on the property at all. One would expect here a query distribution that is completely symmetric with respect to any permutation of the index set $[n]$. Indeed, a candidate for such a canonical scheme is defined as a *sample-based* test in [4]. The sampled-based distribution μ_p corresponds to choosing every index $i \in [n]$ to be queried independently with probability p. Usually p will be $n^{-\gamma}$ for some $1 > \gamma > 0$. It is folly to expect sample-based tests with significantly fewer queries, even for properties with a constant bound on the number of queries for a test, as evidenced already in [2, Proposition 6.9].

In [4] a connection between *proximity-oblivious* testers (POT) as defined in [5] and the sample-based querying scheme was suggested. Proximity oblivious testers are testing algorithms whose querying distribution is the same for any proximity parameter ϵ, where instead the distinguishing probability between inputs in L and inputs ϵ-far from L changes with ϵ (adaptive POT were defined there too). The work [4] showed that for such testers that additionally have the property that all indexes get queried with about the same probability (but not necessarily in an independent manner), there exists a conversion to sample based testers with $p = O(n^{-1/q})$, where the coefficient depends on the distinguishing probability, and the parameter measuring the above-mentioned "probability sameness" of the original test.

For an intuition for the $1/q$ in the exponent, consider a property where the input is composed of "disjoint strings" belonging to $\{0,1\}^q$, where each such string is subject to some restriction (for example, having parity 0). A test with $O(q)$ queries would just sample a few of these strings and check them for their corresponding restrictions. For a sample-based test, we would like to set p so that with high probability some of these strings would be queried in their entirety, and setting it to $O(n^{-1/q})$ does just that.

In [6] it is shown that all 1-sided proximity-oblivious testers over the alphabet $\{0,1\}$ are convertible to the canonical sample-based scheme, where $p = n^{-\gamma}$ with γ depending (somewhat badly) on q, ϵ and the distinguishing probability δ. In [4] there is an example of a testable property that has no sublinear query complexity sample-based test at all, but it works only over an alphabet whose size is exponential in n, and so does not contradict the result of [6].

Here we take the investigation much further, and prove the following.

Theorem 1.1 (informal statement of our main result): Every property of words in Ξ^n, that has a non-adaptive ϵ-test with q queries and detection probability δ (either 1-sided or 2-sided), admits a test using the sample-based canonical querying scheme, where the distribution μ_p has $p = O(n^{-\gamma})$, with γ depending on q, δ, and for 2-sided testing also on $|\Xi|$ and ϵ, and the hidden coefficient depending on q, δ, ϵ and $|\Xi|$.

We prove this separately for 1-sided tests and 2-sided tests. For 2-sided tests we go further and prove the result for *partial tests*, that are only guaranteed to accept inputs in some sub-property L' with high probability, a generalization whose relevance is explained below.

For both 1-sided testing and 2-sided (possibly partial) testing we obtain a very improved bound on γ as compared to the 1-sided testing result of [6]. Additionally, the dependency of the coefficient on $|\Xi|$ is logarithmic, while for the 2-sided test the additional dependency of γ on $|\Xi|$ is of type $\log\log\log(|\Xi|)$. This shows that, for non-adaptive tests, the exponential size of the alphabet in the counter-example in [4] is essential. For non-adaptive tests, we believe that the "correct" γ should be $-1/q$, just like the result in [4] for the more restricted case, but cannot prove it yet.

By the standard conversion of adaptive tests to non-adaptive tests, Theorem 1.1 also holds for adaptive tests. However, in both the 1-sided and the 2-sided cases, the resulting γ parameter is a function of $|\Xi|$ to the power of a polynomial function in q. Note that [4] holds also for restricted adaptive tests with $\gamma = 1/q$, but we are not sure at the moment whether an analogous result should hold for adaptive tests in general.

A. Implications for multitests

There are several motivations for finding canonical testing schemes. One of them is for proving lower bounds, which may be easier when the querying distribution is "simple" and known (for the dense graph model it was

used in [7] to show that a test for triangle-freeness cannot be polynomial in ϵ). Here we would like to highlight another motivation, which also played an implicit role in the original motivation of [6].

Given a sequence of properties $L_1, \ldots, L_r$, a *multitest* for them is an algorithm that makes queries to a word $w \in \Xi^n$, and provides a sequence of answers. With probability at least $1 - \delta$, the answers should be correct for *all* the properties, that is, for every k such that $w \in L_k$ the corresponding answer should be "yes", and for every k such that w is ϵ-far from L_k the corresponding answer should be "no".

If we know nothing else about the properties apart from that they are all testable using q queries each, then the scalability of a test to a multitest would be quasilinear: We first take a test for every L_i, and amplify its success probability to $1 - \delta/r$ (which multiplies the number of queries by $O(\log r)$). Then we just run these r tests one after the other, and use the union bound for the total success probability, all in all using $O(q \cdot r \log r)$ queries. This is not always good enough, as in some applications r can depend on n, and may even be greater than n (say through a polynomial dependency).

However, the situation changes dramatically if we know all properties to share a canonical testing scheme with q' queries (where q' could depend on n). In this case, we can re-use the same queries for all r (amplified) tests, and the union bound will still work. This brings us to using only $O(q' \cdot \log r)$ queries in all. This scalability can have many implications.

In [6], multitests are implicitly used for testing unions of properties. This in turn allows to convert in certain cases tests requiring proofs as per the $\mathcal{MAP}$ scenario (defined in [8] and also developed in [6]) to tests that still have a sublinear query complexity but do not require such proofs. In this setting we deploy the generalization of our result to partial testing, as a $\mathcal{MAP}$ scenario converts to a union of partial testing problems.

Another scenario aided by a multitest is if one wants to store the results for w belonging (approximately) to a rather large set of possible properties. If the properties share a canonical testing scheme, and the corresponding property tests also admit a not too large computation time overhead, then it may be worthwhile to store instead the common set of queries performed by the multitest, because this query set increases rather slowly with r.

Finally, a canonical testing scheme also allows for some measure of privacy: Suppose that one wants to test a particular property of $w \in \Xi^n$, but wants to hide from the "input holder" the identity of the particular property to be tested. By using the canonical testing scheme, no one but the party performing the test can discern which of the properties having the canonical scheme is being tested for.

B. Methods used

The combinatorics of a test: The crucial analysis used for converting a non-adaptive test with q queries to a sample-based test is of a combinatorial nature. We take the support of the query distribution of a non-adaptive test, and analyze it as a family of query sets, essentially a q-uniform hypergraph whose vertex set is the domain of possible queries. We can assume (through a simple processing of the original test) that the number of possible query sets is linear in the domain size n.

For 1-sided tests, since they reduce to checking whether the set of queries is a witness refuting the possibility of the input belonging to the property, the support of the distribution provides most of the information we need. We will mainly analyze the subsets of the hypergraph of possible query sets (considering their forbidden input values) to achieve our result.

For 2-sided tests, since we need a way to estimate (from both directions) the acceptance probability of the original test, the distinctness of the probability values needs to be removed rather than ignored. For this we take the original test and make it into a *combinatorial* one, in the sense that the query distribution will be uniform over the family of possible query sets. Much work is needed to fully convert a general 2-sided test to a combinatorial one that can be analyzed as a hypergraph, and this introduces some extra dependency on the alphabet size. To aid with the analysis of the 2-sided tests, a formalism of *probabilistic formulas* is introduced. The combinatorialization lemma developed here has potential for future uses, as it generalizes to promise problems outside property testing – it is sufficient for the property to have at least one "yes" instance and one "robust no" instance for us to be able to combinatorialize the distinguishing algorithm.

The usefulness of matchings and pompoms: For 1-sided tests, finding large "matchings" (families of disjoint sets) of refuting witnesses would be ideal for sample based testing, but also the assumption of the linear size of our family of sets does not guarantee their existence. For example, it may be the case that there is a query common to all query sets in our family, and so no disjoint sets exist. Another hypothetical case is that all sets are "concentrated" as subsets of a common relatively small set of possible queries. In fact, it is important to the following that this last case can be ruled out for a family of sets that comes from a property test.

For 2-sided tests, a "decomposition" to matchings would have worked well for a combinatorial test, because then the sample based querying would have produced from every matching enough edges for an estimation of the "acceptance ratio" associated with that matching. But if we cannot always find even a single matching, of course we cannot guarantee such a decomposition.

The next option could be finding large sunflowers as defined in [9]. This is the approach taken by [6], and sunflowers can be adapted to deal with a 2-sided test here. However, the obtained γ value for the $n^{-\gamma}$ sampling test would depend very badly on the other parameters, because this approach requires processing in stages.

Here we present a generalization of sunflowers, which we call *pompoms*. The main difference is that the "core" common to the participating sets is not their intersection as in sunflowers, and in fact could be much larger – the only requirements are that the participating sets all have the same intersection size with the core (but not necessarily the same intersection), and are disjoint outside the core. By the method that we outline below, the support of the query distribution is shown to admit pompoms larger than the sunflowers it would admit, and moreover a decomposition of a large portion thereof into pompoms that all share the same core. For a 2-sided (combinatorial) test, we can use this to categorize the possible inputs (inputs in plural, because we will not know anything about what happens within the core itself) in just one processing round.

Constellations in hypergraphs: The pompoms mentioned above are obtained through a general structural result about sparse q-uniform hypergraphs, that may be of independent interest. We find in such hypergraphs a large *constellation* (in fact we prove this also for weighted hypegraphs). A constellation is a set of edges that is guaranteed to have no high degree vertices, outside a small set (which would become the core of our pompoms). In fact we manage to find something stronger, a large *exacting* constellation. The additional requirement that it satisfies is of a small mutual degree condition for all vertex sets of size lesser than q.

For 2-sided tests, we then use the fact that a constellation readily allows for the extraction of pompoms because of the low degree condition. We can extract them one by one greedily to find a decomposition of a large portion of the original hypergraph associated with the test.

For 1-sided tests we can do even better. Since what we obtain is an exacting constellation, we can apply the well-known Janson's inequality directly to prove that we find a witness with high probability. Making sure it happens for every possible assignment to the core, the result of this direct analysis is a better value of γ, the power of n, than what would have been obtained through the extraction of pompoms of witnesses.

II. Preliminaries

A. Large deviation bounds

The following is useful for the analysis of sample based testing.

Lemma 2.1 (multiplicative Chernoff bounds): Let $X_1, \ldots, X_m$ be independent Boolean random variables such that $\Pr[X_i = 1] = p$. Let $X = \sum_{i=1}^{m} X_i$. For any $\gamma \in (0, 1]$,

$$\Pr\left[X > (1+\gamma)pm\right] < \exp\left(-\gamma^2 pm/3\right) \quad \text{and} \quad \Pr\left[X < (1-\gamma)pm\right] < \exp\left(-\gamma^2 pm/2\right)$$

Lemma 2.2 (Hoeffding bounds [10]): Let $Y_1, \ldots, Y_m$ be independent random variables such that $0 \leq Y_i \leq 1$, for $i = 1, \ldots, m$ and let $\eta = \mathrm{E}\left[\sum_{i=1}^{m} Y_i/m\right]$. Then,

$$\Pr\left[\left|\frac{\sum_{i=1}^{m} Y_i}{m} - \eta\right| \geq t\right] \leq 2\exp(-2mt^2)$$

Lemma 2.3 (without replacement [10]): Let $X_1, \ldots, X_m$ be random variables picked uniformly without repetition from the sequence $\mathcal{C} = (\gamma_1, \ldots, \gamma_m)$ where $0 \leq \gamma_i \leq 1$ (this means that $i_1, \ldots, i_m$ are picked uniformly without repetition from $[m]$, and then every X_j is set to γ_{i_j}; it may be that some γ_i are equal to others). Let $Y_1, \ldots, Y_m$ be independent random variables picked with repetition from $\mathcal{C}$ (i.e. every k_j is uniformly and independently chosen from $[m]$ and then Y_j is set to γ_{k_j}). Then the conclusion of Lemma 2.2 for $Y_1, \ldots, Y_m$ holds also for $X_1, \ldots, X_m$, that is, $\Pr\left[|\sum_{i=1}^m X_i/m - \eta| \geq t\right] \leq 2\exp(-2mt^2)$ where $\eta = \frac{1}{m}\sum_{i=1}^m \gamma_i$.

Lemma 2.4 (large deviation bound): Denote by μ_p the distribution over subsets of $[m]$, where every $i \in [m]$ is picked into the subset with probability exactly p, independently from all other $j \neq i$. Suppose that $\gamma_1, \ldots, \gamma_m$ are all values in $[0, 1]$, and let $U \subseteq [m]$ be chosen according to μ_p, where $p \geq 10c/\eta^2 m$ and $c > 1$. Then, with probability at least $1 - e^{-c}$, the value $(\sum_{i\in U} \gamma_i)/|U|$ (where we arbitrarily set it to $\frac{1}{2}$ if $U = \emptyset$) is in the range $(\sum_{i=1}^m \gamma_i)/m \pm \eta$.

Proof: If U is picked according to μ_p, then $\mathrm{E}[|U|] = pm$. By the multiplicative Chernoff bound of Lemma 2.1 we have the following bound on the probability of the size of U being small:

$$\Pr\left[|U| \leq \frac{pm}{2}\right] \leq \exp(-pm/8) \leq e^{-c}/2.$$

We continue our analysis conditioned on the event that the size of U is at least $pm/2$. For every $k \geq pm/2$, let us analyze separately the deviation of the value $\left(\sum_{i\in U} \gamma_i\right)/|U|$ conditioned on $|U| = k$. Lemma 2.3 holds for this case, stating that the probability of $\left|\frac{1}{k}\sum_{i=1}^k X_i - \frac{1}{m}\sum_{i=1}^m \gamma_i\right|$ being greater than η is bounded by $e^{-c}/2$. Hence, for U picked according to μ_p, the probability of $\sum_{i\in U} \gamma_i/|U|$ being outside the range $\sum_{i\in m} \gamma_i/m \pm \eta$ is at most e^{-c}, by the union bound on the event of $|U| < pm/2$, and the event of $\left|\frac{1}{k}\sum_{i=1}^k X_i - \frac{1}{m}\sum_{i=1}^m \gamma_i\right| > \eta$ while $|U| \geq pm/2$. ∎

We conclude this section with Janson's inequality, that in some cases can be used directly to show that a universal sampling scheme is likely to capture at least one of a family of sets in its entirety.

Lemma 2.5 (Janson's inequality, [11]): Let R be a random subset of $\{1, \ldots, n\}$ such that $\Pr[i \in R] = p$. Let $\{Q_i\}_{i\in I}$ be subsets of $\{1, \ldots, n\}$ where I is an index set. Let E_i be the event that a set $Q_i \subseteq R$. Define $\Delta = \sum_{i\neq j, Q_i\cap Q_j\neq\emptyset} \Pr[E_i \wedge E_j]$ and set $\eta = \sum_{i\in I} \Pr[E_i]$. Then,

$$\Pr\left[\bigwedge_{i\in I} \neg E_i\right] \leq e^{-\eta+\Delta/2}. \tag{1}$$

If $\Delta \geq \eta$, then

$$\Pr\left[\bigwedge_{i\in I} \neg E_i\right] \leq e^{-\eta^2/2\Delta}. \tag{2}$$

B. Words and distributions

Notation for words: Let Ξ be an alphabet, and let $w \in \Xi^n$, $i \in [n]$ and $Q \subseteq [n]$. We use w_i to denote the i'th letter of w and w_Q to denote the word $v \in \Xi^{|Q|}$ such that, for every $j \in [|Q|]$, $v_j = w_{Q(j)}$, where $Q(j)$ is the j'th smallest member of Q. Let $C \subseteq [n]$ and σ be a word in $\Xi^{|C|}$. We denote by $w_{\sigma,C}$ the word that we get by taking w and replacing its sub-word w_C with σ.

Definition 2.6 (word distances): Two words $w, v \in \Xi^n$ are said to be *ϵ-far* if there is no A of size at most ϵn for which $w_{[n]\setminus A} = v_{[n]\setminus A}$ (in other words, we use the normalized Hamming distance). Otherwise these words are said to be *ϵ-close*. Given a property $L \subseteq \Xi^n$, a word w is said to be *ϵ-close* to L if there exists an ϵ-close word v which is in L, and otherwise w is said to be *ϵ-far* from L.

Notation for distributions: We deal with distributions μ over subsets of $[n]$. For $A \subseteq [n]$ we denote by $\mu(A)$ the probability of A being drawn by μ. For a non-empty event, that is a family of sets $\emptyset \neq \mathcal{A} \subseteq 2^{[n]}$, we abuse notation somewhat and denote $\mu(\mathcal{A}) = \sum_{A\in\mathcal{A}} \mu(A)$. We denote by $\mathrm{Supp}(\mu)$ the family of positive probability outcomes $\{A \subseteq [n] : \mu(A) > 0\}$, and for two distributions μ and μ' denote by $\mathrm{dist}(\mu, \mu')$ the variation distance $\frac{1}{2}\sum_{A\subseteq[n]} |\mu(A) - \mu'(A)| = \max_{\mathcal{A}\subseteq 2^{[n]}} |\mu(\mathcal{A}) - \mu'(\mathcal{A})|$.

C. Property Testing

We start this subsection by defining tests. We define partial tests (of which tests are a special case), because we would like our main result to also have applications in the realm of $\mathcal{MAP}$s as defined in [8].

Definition 2.7 (partial (ϵ, δ, q)-test): Given two properties $L' \subseteq L \subseteq \Xi^n$, a partial (ϵ, δ, q)-*test* for (L', L) is a randomized algorithm $\mathcal{A}$ that, given query access to the input w, uses q queries and satisfies the following:

1) If $w \in L'$, then $\Pr[\mathcal{A}(w) = 1] \geq 1 - \delta$.
2) If w is ϵ-far from L, then $\Pr[\mathcal{A}(w) = 0] \geq 1 - \delta$.

The test is 1-*sided* if, when $w \in L'$, the output is always 1, and otherwise it is 2-*sided*. If the choice of every query is independent of the answers to the previous queries, then the test is *non-adaptive*, and otherwise it is *adaptive*. In the case where $L' = L$, we call it an (ϵ, δ, q)-*test* for L.

We remark that, in the case of a non-adaptive test, we may assume that the set of queries is selected before any query is made. So, a non-adaptive test can be viewed as consisting of three steps: (i) a set of queries Q is randomly selected according to a distribution over $2^{[n]}$; (ii) the sub-word w_Q is queried; (iii) the output is computed according to (Q, w_Q).

We note that a 1-sided test can reject only if (Q, w_Q) constitutes a proof that w is not in the property. This occurs if and only if Q is a *witness against* w as defined next.

Definition 2.8 (witness against a word): A set $Q \subseteq [n]$ is a *witness* against a word $w \in \Xi^n$ (with regards to a property L), if every $u \in \Xi^n$ such that $u_Q = w_Q$ is not in L.

Without loss of generality, we assume that a test always rejects when it encounters a witness against the word. In the case of a 1-sided tester it is actually the case that the test rejects only if it encountered a witness. We next formally define the concept of the distribution of a non-adaptive test.

Definition 2.9 (distribution of a non-adaptive (ϵ, δ)-test): The distribution of a non-adaptive (ϵ, δ)-test $\mathcal{A}$, denoted by $\mu_{\mathcal{A}}$, is a distribution over $2^{[n]}$, such that, for every $Q \subseteq [n]$, the value of $\mu_{\mathcal{A}}(Q)$ is the probability that $\mathcal{A}$ will select Q to be its set of queries. We omit the subscript when it is clear from context.

Our conversion results rely on the combinatorial aspects of distributions of tests. In fact, for non-adaptive 1-sided tests, without loss of generality this distribution is the sole defining object, because the test can be assumed to reject if and only if its query set produced a witness against the input word. In particular, we show a reduction to the case where the cardinality of the support of the distribution has a bound linear in n. We use the following definition to capture this case and afterwards we give the reduction.

Definition 2.10 (non-adaptive (ϵ, δ, q, k)-test): A non-adaptive (ϵ, δ, q)-test (or partial test) is an (ϵ, δ, q, k)-test, if $|\operatorname{Supp}(\mu)| \leq k$.

We observe that the support of the distribution of an (ϵ, δ, q)-test contains only sets of cardinality q. We use the term (ϵ, δ)-*test* (omitting q) when we do not make any assumption on the cardinality of the sets in the distribution. The following lemma transforms a 1-sided test to one with parameters more suitable for analysis and conversion to sample-based testing.

Lemma 2.11: A non-adaptive 1-sided $(\epsilon/2, \delta, q)$-test can be converted to a non-adaptive 1-sided $(\epsilon/2, 1/2(q' + 1), q', 4(q' + 1)\log(|\Xi|)n)$-test where $q' = O(q\log(q)/(1 - \delta))$.

Proof: First, by traditional amplification, repeating the original test $20\log(q)/(1 - \delta)$ times and rejecting if any run had rejected, we convert it to an $(\epsilon/2, 1/1000q'', q'')$-test where $q'' = O(q\log(q)/(1 - \delta))$. Then we consider the outcome of running the test $10\log(|\Xi|)q''n$ times independently. By Lemma 2.1, for any fixed $\epsilon/2$-far input $w \in \Xi^n$, the probability that it is accepted by more than a $1/10q''$ fraction of the runs is bounded by $e^{-99^2 \cdot \log(|\Xi|)n/3000} < \frac{1}{2}|\Xi|^{-n}$. This means that with probability at least $\frac{1}{2}$, such a sequence of runs will satisfy the above for all $\epsilon/2$-far inputs at once. We fix such a sequence of runs, and make it the new test. That is, the new μ' consists of selecting one of the fixed runs uniformly at random, and using its query set. This brings us to an $(\epsilon/2, 1/10q'', q'', 10\log(|\Xi|)q''n)$-test. We artificially increase the number of queries to $q' = 3q''$ to obtain our required test. ∎

The following two lemmas are essential for the analysis of 1-sided tests. When reading them, one should keep in mind that, when they are applied, the δ parameter in their statement is small since the tests analyzed are those implied by Lemma 2.11.

Lemma 2.12: Let $\mathcal{J} \subseteq \mathrm{Supp}(\mu)$, where μ is the distribution of a 1-sided $(\epsilon/2, \delta)$-test for a non-empty property L for which ϵ-far words exist. If $|\bigcup_{Q\in\mathcal{J}} Q| < \epsilon n/2$, then $\mu(\mathcal{J}) < \delta$.

Proof: Let $T = \bigcup_{Q\in\mathcal{J}} Q$, let $u \in \Xi^n$ be a word in L and $w \in \Xi^n$ be ϵ-far from L, and let $v = w_{u_T,T} \in \Xi^n$ be such that $v_T = u_T$ and $v_{[n]\setminus T} = w_{[n]\setminus T}$. Assume that $|\bigcup_{Q\in\mathcal{J}} Q| < \epsilon n/2$. Then, by the triangle inequality, v is $\epsilon/2$-far from L.

Considering a 1-sided test of v with distribution μ, we first note that no member of $\mathcal{J}$ is a witness against v. Thus, $\mu(\mathcal{J})$ is at most 1 minus the probability of μ obtaining a witness. As v is $\epsilon/2$-far from L, the probability of obtaining a witness by μ is at least $1 - \delta$, implying that $\mu(\mathcal{J}) < \delta$. ■

Lemma 2.13: For μ which is the distribution of a 1-sided $(\epsilon/2, \delta)$-test for a non-empty property L for which ϵ-far words exist, let $w \in \Xi^n$ be $5\epsilon/6$-far from L and $\mathcal{J} \subseteq \mathrm{Supp}(\mu)$. If $\mu(\mathcal{J}) \geq 2\delta$, then the set $\mathcal{S}$ of all $Q \in \mathcal{J}$ which are witnesses against w satisfies $|\bigcup_{Q\in\mathcal{S}} Q| \geq \epsilon n/2$.

Proof: Let $\mathcal{S} \subseteq \mathcal{J}$ be the subset of witnesses against w as in the formulation of the lemma. Since w is $5\epsilon/6$-far from L, the distribution μ provides a witness against w with probability at least $1 - \delta$, and therefore $\mu(\mathcal{S}) \geq \delta$. Consequently, by Lemma 2.12, $|\bigcup_{Q\in\mathcal{J}} Q| \geq \epsilon n/2$. ■

Definition 2.14 (p-sampling (ϵ, δ)-test): A *p-sampling test* for a property L is an (ϵ, δ)-test such that every $i \in [n]$ is selected as a query, independently, with probability p; in other words, it is a sample-based test with probability p as defined in [4]. A p-sampling test is 1-sided if every word in the property is accepted with probability 1 and otherwise it is 2-sided.

We use the notation μ_p to denote the distribution of the query sets of the p-sampling test. In other words, for every $A \subseteq [n]$ we define $\mu_p(A) = p^{|A|}(1-p)^{n-|A|}$.

III. Hypergraph structure, constellations and pompoms

Our analysis of non-adaptive property tests has a strong combinatorial component. Generally, a non-adaptive test with q queries is correlated with a family $\mathcal{Q}$ of subsets of $[n]$, namely the support of its query set distribution. Since we assume that the test always makes exactly q queries, all sets in this family are of size q, making it a q-uniform hypergraph.

In this section we provide a strong structural theorem for such families that are sparse, i.e. families whose size is linear in n. Essentially, unless a large portion of the family is "concentrated" over very few vertices, we can find a large sub-family where all vertex sets have the "expected" bound on their degrees (here the degree of a vertex set is the number of the members of $\mathcal{Q}$ that contain it; the special case where the set is of size 1 corresponds to the usual definition of vertex degrees). The structure that we find in $\mathcal{Q}$ is of the following type.

Definition 3.1 (constellation and exacting constellation): Given a family $\mathcal{Q}$ of subsets of size q of $[n]$, for $i \in [q]$ and any positive number η, an *(η, i)-constellation* is a pair $(\mathcal{S}, C)$ consisting of a set $C \subseteq [n]$ and a family $\mathcal{S} \subseteq \mathcal{Q}$ satisfying the following.

1) $|C| \leq \eta n^{1-i/q}$.
2) $|Q \setminus C| = i$, for every $Q \in \mathcal{S}$.
3) If $i > 1$, then every $j \in \bigcup_{Q\in\mathcal{S}} Q \setminus C$ is in at most $n^{(i-1)/q}$ sets from $\mathcal{S}$.

The constellation is *exacting* if the following stronger condition is satisfied.

4) If $i > 1$, then every $J \subseteq \bigcup_{Q\in\mathcal{S}} Q \setminus C$ of size less than i is contained in at most $n^{(i-|J|)/q}$ sets from $\mathcal{S}$.

For our analysis it will be convenient to state a version for a weighted hypergraph, i.e. a family of sets $\mathcal{Q}$ supplied with a distribution μ over it. An unweighted (enumerative) version would then be obtained as the special case where μ is the uniform distribution over $\mathcal{Q}$. Our main structural result is the following.

Lemma 3.2 (exacting constellations in sparse unconcentrated hypergraphs): Suppose that $\mathcal{Q}$ is a family of subsets of size q of $[n]$, and suppose that μ is a distribution over $\mathcal{Q}$. Furthermore suppose that $|\mathcal{Q}| \leq \eta n$,

and denoting by C_0 the set of indexes $j \in [n]$ contained in at least $n^{1/q}$ sets from $\mathcal{Q}$ and by $\mathcal{S}_0$ the family of members of $\mathcal{Q}$ contained in C_0, suppose that $\mu(\mathcal{S}_0) \leq \frac{1}{q+1}$. Under these conditions there exists $i \in [q]$ so that $\mathcal{Q}$ contains an exacting (η, i)-constellation $(\mathcal{S}, C)$ for which $\mu(\mathcal{S}) \geq \frac{1}{q+1}$.

Noting that in the above statement C_0 is necessarily of size not more than $\eta n^{1-1/q}$, we note that the condition on $\mu(\mathcal{S}_0)$ means that not too much of the family $\mathcal{Q}$ is concentrated into this small set.

Before we move to prove Lemma 3.2 above, let us observe the usefulness of constellations. As will be explained in the following sections, sample based testing hypothetically works best when used over *matchings*, families of disjoint query sets. Failing that, one would like to work with families that are at least guaranteed to be disjoint outside a small "core" set C. This is embodied in the following definition.

Definition 3.3: (i-pompom) A family of sets $\mathcal{S}$ is an *i-pompom* if there exists a set C, which we refer to as the *core* of the i-pompom, such that the following hold.

1) $|Q \setminus C| = i$ for every $Q \in \mathcal{S}$.
2) $Q \setminus C$ and $Q' \setminus C$ are pairwise disjoint for every distinct Q and Q' in $\mathcal{S}$.

The restriction of the cardinality of the sets $Q \setminus C$ to be the same for every $Q \in \mathcal{S}$ is required to support technical computations in later proofs.

Once we have a constellation (even if not an exacting one) whose sets cover sufficiently many vertices, it is east to "pull" a corresponding pompom out of it.

Observation 3.4: Suppose that $(\mathcal{S}, C)$ is an (η, i)-constellation for some η, and that $|\bigcup_{Q\in\mathcal{S}} Q| - |C| \geq \beta n$ for some β. Then there exists $\mathcal{S}' \subseteq \mathcal{S}$ which is an i-pompom with core C, whose size is at least $(\beta/i) \cdot n^{1-(i-1)/q}$.

Proof: Let $\mathcal{B}$ be the family $\{Q \setminus C : Q \in \mathcal{S}\}$, and observe that $|\bigcup_{R\in\mathcal{B}} R| \geq |\bigcup_{Q\in\mathcal{S}} Q| - |C| \geq \beta n$.

Suppose that $i = 1$. We let $\mathcal{S}' \subseteq \mathcal{S}$ be maximal so that, for every $Q \in \mathcal{S}'$, $Q \setminus C$ is distinct. Clearly, $|\mathcal{S}'| \geq \beta \cdot n$, and $\mathcal{S}'$ is a 1-pompom with core C.

Suppose now that $i > 1$. Then there exists $\mathcal{B}' \subseteq \mathcal{B}$ such that every pair of sets in $\mathcal{B}'$ is disjoint and $|\mathcal{B}'| \geq (\beta/i) \cdot n^{1-(i-1)/q}$, because every $j \in \bigcup_{R\in\mathcal{B}} R = \bigcup_{Q\in\mathcal{S}} Q \setminus C$ is in at most $n^{(i-1)/q}$ sets from $\mathcal{S}$ and hence from $\mathcal{B}$ (so we can just pick the members of $\mathcal{B}'$ greedily). We let $\mathcal{S}' \subseteq \mathcal{S}$ be maximal so that, for every $Q \in \mathcal{S}'$, $Q \setminus C$ is a distinct member of $\mathcal{B}'$. Clearly, $|\mathcal{S}'| = |\mathcal{B}'| \geq (\beta/i) \cdot n^{1-(i-1)/q}$, and $\mathcal{S}'$ is an i-pompom with core C. ■

We also note the following trivial observation.

Observation 3.5: If $(\mathcal{S}, C)$ is an (η, i)-constellation, then so is $(\mathcal{T}, C)$ for any $\mathcal{T} \subseteq \mathcal{S}$.

We now turn to prove Lemma 3.2. We start by defining three sequences, $\{\mathcal{M}_i\}_{i=0}^q$, $\{C_i\}_{i=0}^q$ and $\{\mathcal{S}_i\}_{i=0}^q$, where we will prove that $\{\mathcal{S}_i\}_{i=0}^q$ is a partition of $\mathcal{Q}$. We will also prove that for every $i \in [q]$, the sets $\mathcal{S}_i$ and C_i respectively compose an exacting (η, i)-constellation.

Definition 3.6 ($\mathcal{M}_i$, C_i and $\mathcal{S}_i$): Given a family $\mathcal{Q}$ of subsets of size q of $[n]$, whose cardinality is bounded by ηn, we inductively define $\mathcal{M}_i$, C_i and $\mathcal{S}_i$ as follows.

1) Let $\mathcal{M}_0 = \mathcal{Q}$, and C_0 be the set of indexes $j \in [n]$ such that j is a member of at least $n^{q^{-1}}$ sets in $\mathcal{M}_0$.
2) For $i = 0, 1, \ldots, q$, after $\mathcal{M}_i$ and C_i are defined, let $\mathcal{S}_i$ be the family of all sets $Q \in \mathcal{M}_i$ such that $|Q \cap C_i| = q - i$.
3) For $i = 1, \ldots, q$, after $\mathcal{M}_i$ is defined, let C_i be the set of indexes $j \in [n]$ such that j is a member of at least $n^{i/q}$ sets in $\mathcal{M}_i$.
4) For $i = 1, \ldots, q$, after $\mathcal{M}_{i-1}$ and $\mathcal{S}_{i-1}$ are defined, let $\mathcal{M}_i = \mathcal{M}_{i-1} \setminus \mathcal{S}_{i-1}$.

Before we continue, let us note that C_0 and $\mathcal{S}_0$ are exactly the same as the corresponding sets appearing in the statement of Lemma 3.2. The following statement gives the properties of these sequences.

Observation 3.7: The following hold for the sets of Definition 3.6 when they are constructed from a family $\mathcal{Q}$ satisfying the conditions there.

1) $|C_0| \leq \eta q n^{1-1/q}$.
2) $|C_i| \leq \eta q n^{1-i/q}$ for all $1 \leq i \leq q$.
3) $\mathcal{M}_i \subseteq \mathcal{M}_{i-1}$ for all $1 \leq i \leq q$.
4) $C_i \subseteq C_{i-1}$ for all $1 \leq i \leq q$.

5) $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for all $1 \leq i < j \leq q$.

Proof: Item 3 follows immediately from the construction, and implies Item 4 in turn. Item 1 follows from the construction along with the assumption on the cardinality of $\mathcal{Q}$, and so does Item 2 using Item 3. For Item 5 assume that $i < j$, and note the construction of $\mathcal{M}_{i+1}$, which makes it disjoint from $\mathcal{S}_i$, while containing $\mathcal{S}_j$ by Item 3. ∎

The goal of the following two lemmas is to prove that, for every $i \in [q]$ and $Q \in \mathcal{S}_i$, we have that $(Q \setminus C_i) \cap C_{i-1} = \emptyset$. According to Definition 3.6, this implies that C_i and $\mathcal{S}_i$ satisfy Condition 3 of Definition 3.1. At a high level of abstraction the proof starts with the assumption that for some $i \in [q]$ there exists $Q \in \mathcal{S}_i$ such that $(Q \setminus C_i) \cap C_{i-1} \neq \emptyset$; afterwards it is shown that this Q is in $\mathcal{M}_j$ for some $j \in [i-1]$; this by Definition 3.6 implies that $Q \notin \mathcal{S}_i$ in contradiction to the assumption that $Q \in \mathcal{S}_i$.

The following lemma is used to restrict the setting to the case where $j = i - 1$, also showing that $\mathcal{S}_0, \ldots, \mathcal{S}_q$ is a partition of $\mathcal{Q}$.

Lemma 3.8: For $i = 0, 1, \ldots, q$ and every $Q \in \mathcal{M}_i$ we have that $|Q \cap C_i| \leq q - i$; in particular $\mathcal{S}_q = \mathcal{M}_q$ and thus $\mathcal{S}_0, \ldots, \mathcal{S}_q$ partition $\mathcal{Q}$.

Proof: By definition, $|Q| \leq q$ for every $Q \in \mathcal{M}_0$. Hence, $|Q \cap C_0| \leq q - 0 = q$. We proceed by induction over i. Assume that the statement of the lemma holds for $i - 1 \geq 0$. Suppose for the sake of contradiction that there exists $Q \in \mathcal{M}_i$ such that $|Q \cap C_i| > q - i$. Since $C_i \subseteq C_{i-1}$ by Item 4 of Observation 3.7, this implies that $|Q \cap C_{i-1}| \geq q - (i-1)$. Hence, $|Q \cap C_{i-1}| = q - (i-1)$, by the induction assumption. Therefore, $Q \in \mathcal{S}_{i-1}$, because, by construction, we also have that $Q \in \mathcal{M}_{i-1}$. Consequently, we get the contradiction that $Q \notin \mathcal{M}_i$, since $\mathcal{M}_i = \mathcal{M}_{i-1} \setminus \mathcal{S}_{i-1}$.

Having proved the first part of the lemma, it implies that $|Q \cap C_q| = 0$ for every $Q \in \mathcal{M}_q$, and hence by Item 2 of Definition 3.6 we have $\mathcal{S}_q = \mathcal{M}_q$. ∎

Lemma 3.9: For $i = 1, \ldots, q$ and every $Q \in \mathcal{S}_i$ we have that $(Q \setminus C_i) \cap C_{i-1} = \emptyset$.

Proof: We proceed by induction over i. The base case is $i = 0$ which follows from the definition of $\mathcal{S}_0$, even if we set $C_{-1} = [n]$. Assume that the statement of the lemma holds for $i - 1 \geq 0$. Suppose for the sake of contradiction that there exists $Q \in \mathcal{S}_i$ such that $|(Q \setminus C_i) \cap C_{i-1}| > 0$. Thus, $|Q \cap C_{i-1}| = |(Q \setminus C_i) \cap C_{i-1}| + |Q \cap C_i| \geq q - (i-1)$, because $|Q \cap C_i| = q - i$ by Item 2 of Definition 3.6, and $C_i \subseteq C_{i-1}$ by Item 4 of Observation 3.7. Therefore, by Lemma 3.8, $|Q \cap C_{i-1}| = q - (i-1)$. Since by construction we also have that $Q \in \mathcal{M}_{i-1}$ we deduce that $Q \in \mathcal{S}_{i-1}$. Consequently, we get the contradiction that $Q \notin \mathcal{S}_i$, since $\mathcal{S}_i \subseteq \mathcal{M}_i = \mathcal{M}_{i-1} \setminus \mathcal{S}_{i-1}$. ∎

At this point we can also prove that C_i and $\mathcal{S}_i$ also satisfy the stronger Condition 4 of Definition 3.1, relating to exact constellations.

Lemma 3.10: For every $1 \leq j < i \leq q$ and any set J disjoint from C_i for which $|J| = j$, we have that $\mathcal{S}_i$ contains no more than $n^{(i-j)/q}$ sets containing J.

Proof: Supposing otherwise, let J be a set violating the assertion of the lemma. This means that all members of J are inside C_{i-j}, since $\mathcal{S}_i \subseteq \mathcal{M}_i \subseteq \mathcal{M}_{i-j}$ by Observation 3.7. Since also $C_i \subseteq C_{i-j}$ by Observation 3.7, it would mean that every member Q of $\mathcal{S}_i$ that contains J would satisfy $|Q \cap C_{i-j}| \geq (q-i) + j = q - (i-j)$. On the other hand, by Lemma 3.8 we have that $|Q \cap C_{i-j}| \leq q - (i-j)$, and so we have equality there, meaning that Q should have been put in $\mathcal{S}_{i-j}$, a contradiction. ∎

Now we can sum up the lemmas about these defined sequences.

Lemma 3.11: For every $i \in [q]$ the pair $(\mathcal{S}_i, C_i)$ is an exact $(\eta q, i)$-constellation.

Proof: The three conditions for a constellation follow from Observation 3.7 Items 1 and 2, Definition 3.6 Item 2, and Lemma 3.9 together with Definition 3.6 Items 1 and 3, in that order. For the condition making this an exact constellation we use Lemma 3.10. ∎

And we can finalize.

Proof of Lemma 3.2: Following from the assumption that $\mu(\mathcal{S}_0) \leq 1/(q+1)$, Lemma 3.8, and Item 5 of Observation 3.7, by averaging, there exists $i \in [q]$ such that $\mu(\mathcal{S}_i) \geq 1/(q+1)$. For this i we pick the pair $(\mathcal{S}_i, C_i)$, which by Lemma 3.11 is the required exact $(\eta q, i)$-constellation. ∎

IV. A CONVERSION OF A 1-SIDED TEST TO A 1-SIDED SAMPLING TEST

We show here that if a property is testable with a 1-sided error, then it has a p-sampling 1-sided (ϵ, δ)-test with p corresponding to some negative power of n. Specifically, we prove the following theorem, which as we explain immediately afterwards implies our claimed result.

Theorem 4.1: For $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$, if a property over Ξ^n has a 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test, then it has a p-sampling 1-sided $(\epsilon, 1/2)$-test where $p = 15\ln|\Xi| \cdot q^4(q+1)^2 n^{-1/q}/\epsilon$.

The preceding theorem is effective for all properties with 1-sided $(\epsilon/2, \delta, q)$-tests, since, by Lemma 2.11, an $(\epsilon/2, \delta, q)$-test can be converted to an $(\epsilon/2, 1/2(q'+1), q', 4(q'+1)\log(|\Xi|)n)$-test, where q' is bounded by a polynomial in q and $1/(1-\delta)$.

We next sketch a proof that the statement of Theorem 4.1 holds for every test that satisfies the additional constraint that it has a distribution μ such that $\mathrm{Supp}(\mu)$ consists of pairwise disjoint sets. The main result of this section is inspired by the idea of a reduction to this simplistic case.

Suppose that μ is a distribution over disjoint query sets of a 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test for L. Let w be ϵ-far from a property L, and $\mathcal{B}$ be the family of all the sets in $\mathrm{Supp}(\mu)$ that are witnesses against w. Now note that if $|\mathcal{B}|$ is sufficiently large, then using the fact that these sets are pairwise disjoint it is easy to show that, with probability at least $1/2$, the set of queries used by a p-sampling test contains at least one of these sets. This in turn implies that a p-sampling test rejects w with probability at least $1/2$. We next explain why $|\mathcal{B}|$ is indeed sufficiently large. Let B be the union of all the sets in $\mathcal{B}$. By definition, the test rejects w with probability at least $1/2$, and therefore $\mu(\mathcal{B}) \geq 1/2$. Thus, by Lemma 2.12, $|B| \geq \epsilon n/2$ and hence $|\mathcal{B}| \geq \epsilon n/(2q)$.

When the sets in $\mathrm{Supp}(\mu)$ are not pairwise disjoint the preceding idea does not work, for example in the case where the size of the intersection of all the sets in $\mathrm{Supp}(\mu)$ is a set $\{i\}$ of size 1. Here, with probability $(1-p)$, a set of queries selected at random according to μ_p will not contain $\{i\}$, and hence will not contain any set from $\mathrm{Supp}(\mu)$ that could be a witness against w. Thus, we have no indication that a p-sampling test rejects w with the required probability. We now explain how to circumvent this barrier in two steps: in the first step we assume that C, a set containing all intersections of sets in $\mathrm{Supp}(\mu)$, is significantly smaller then $\epsilon n/2$, and that we are given w_C in advance; in the second step we show what to do when w_C is not known in advance.

Let $\mathcal{B}$ and B be as defined in the simplistic disjoint query sets case. In the same manner as the simplistic case, we can conclude that $|B| \geq \epsilon n/2$. Let $\mathcal{M}$ be the family of all non-empty sets $Q \setminus C$ such that $Q \in \mathcal{B}$. We note that the size of $\mathcal{M}$ is $O(\epsilon n/q)$ since, by construction, the size of the union of these sets is $|B| - |C| = O(\epsilon n)$. It is thus easy to show that, with high probability, a set of queries selected at random according to μ_p together with C contains a witness against w, and hence a p-sampling test, with the advance knowledge of w_C, rejects w with the required probability. The (exacting) constellation structure that we will eventually utilize here allows also for some ("lesser") intersections outside C itself, so a witness will be found using Janson's bound, rather than the above analysis for completely disjoint sets.

Now we take care of us not knowing w_C. If $(\mathcal{S}, C)$ is a large exact constellation, then for any w that is far from the property, and for every $\sigma \in \Xi^{|C|}$, then there are many witness sets for $w_{\sigma,C}$ in the constellation. We show that in this case the set sampled according to μ_p contains a witness for $w_{\sigma,C}$ with a high enough probability for us to use a union bound over all possible σ. Thus we will obtain a single query set disproving all possible σ at once, a set which we will refer to as a *super-witness*.

Definition 4.2 (super-witness against a word): We say that $X \subseteq [n]$ is a *super-witness* against a word $w \in \Xi^n$, if there exists a set $Y \subseteq [n] \setminus X$ such that, for every $\sigma \in \Xi^{|Y|}$, there exists a set $Q \subseteq X \cup Y$ which is a witness against $w_{\sigma,Y}$.

Recall that the set X in the above definition does not necessarily contain any set from $\mathrm{Supp}(\mu)$. However, as we prove next, it is sufficient to imply that w in not in the property.

Observation 4.3: Any set containing a witness against a word w is also a witness against it. Additionally, a set is a super-witness against w if and only if it is a witness against it.

Proof: The part about containing sets follows immediately from the definition. Additionally, a witness X is also a super-witness by setting $Y = \emptyset$. Now let $X \subseteq [n]$ be a super-witness against $w \in \Xi^n$. By the definition of

a super-witness, for every $u \in \Xi^n$ such that $u_X = w_X$, there exists a witness against u (some subset of $X \cup Y$) and hence $u \notin L$. Thus X is a witness against w with regards to L. ∎

We now prove the main lemma of this section, which shows that if the property L has a 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test admitting an exact $(4q(q+1)\log(|\Xi|), i)$-constellation for some $i \in [q]$, then the $\alpha n^{-1/q}$-sampling algorithm provides an $(\epsilon, 1/2)$-test for the property.

Lemma 4.4: Let $i \in [q]$, $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$, $\alpha = 15\ln|\Xi| \cdot q^4(q+1)^2/\epsilon$, w be any word in Ξ^n that is ϵ-far from L, and μ be the distribution of a 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test for L. If there exists an exact $(4q(q+1)\log(|\Xi|), i)$-constellation for μ where $i \in [q]$, then the $\alpha n^{-1/q}$-sampling algorithm rejects w, with probability at least $\frac{1}{2}$.

Proof: Let $(\mathcal{S}, C)$ be an exact $(4q(q+1)\log(|\Xi|), i)$-constellation. Let $w \in \Xi^n$ be a word that is ϵ-far and $\sigma \in \Xi^{|C|}$. Let $\mathcal{S}_\sigma$ be the subset of $\mathcal{S}$ which contains the set of witnesses against $w_{\sigma,C}$ and let $\mathcal{B}_\sigma$ be the set $\{Q \setminus C \mid Q \in \mathcal{S}_\sigma\}$. Note that for $Q_i \neq Q_j$ in S_σ it is possible that $Q_i \setminus C = Q_j \setminus C$, but we will argue that the sets in $\mathcal{B}_\sigma$ cover many points, and hence there are many distinct sets in $\mathcal{B}_\sigma$. We will prove that an $\alpha n^{-1/q}$-sampling tester fails to sample a set in $\mathcal{B}_\sigma$ with probability at most $\frac{1}{2}|\Xi|^{-|C|}$. The lemma will then follow from a union bound over all strings in $\Xi^{|C|}$.

Since $n > (24q(q+1)^2(\log(|\Xi|))^2/\epsilon)^q$, $|C| < \epsilon n/6$. In particular, by the triangle inequality $w_{\sigma,C}$ is $5\epsilon/6$-far from L. Furthermore, since $\mu(S) \geq \frac{1}{q+1}$, by Lemma 2.13, we obtain $|\bigcup_{Q_j \in \mathcal{S}_\sigma} Q_j| \geq \epsilon n/2$. Therefore we also have $|\bigcup_{Q_j \in \mathcal{B}_\sigma} Q_j| \geq \epsilon n/3$. Combining this with the fact that $|Q_j| = i$ for every $Q_j \in \mathcal{B}_\sigma$, we observe that $|\mathcal{B}_\sigma| \geq \epsilon n/3i$.

For a set $Q_j \in \mathcal{B}_\sigma$, let E_j denote the event that the set Q_j is chosen by the $\alpha n^{-1/q}$-sampler. Let $\eta = \sum_{Q_j \in \mathcal{B}_\sigma} \Pr[E_j] = |B_\sigma|(\alpha n^{-1/q})^i \geq \epsilon n(\alpha n^{-1/q})^i/3i$, and $\Delta = \sum_{j \neq k, Q_j \cap Q_k \neq \emptyset} \Pr[E_j \wedge E_k]$.

Since $(\mathcal{S}, C)$ is an exact $(4q(q+1)\log(|\Xi|), i)$-constellation, we know that the number of sets in $\mathcal{S}$ that have an intersection with Q_j of size r is at most $n^{(i-r)/q}$. This gives us the following upper bound for Δ:

$$\Delta \leq |\mathcal{B}_\sigma| \sum_{r=1}^{i-1} \binom{i}{r} n^{(i-r)/q} (\alpha n^{-1/q})^{2i-r} = |\mathcal{B}_\sigma| \sum_{r=1}^{i-1} \binom{i}{r} n^{-i/q} \alpha^{2i-r} = |\mathcal{B}_\sigma| n^{-i/q} \cdot \alpha^i \sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r}$$

Now, if $\Delta < \eta$, then $\Pr\left[\bigwedge_{Q_j \in \mathcal{B}_\sigma} \neg E_j\right] \leq e^{-\eta + \Delta/2} \leq e^{-\eta/2}$ by Equation 1 in Theorem 2.5. From the bound of η, we get that $\Pr\left[\bigwedge_{Q_j \in \mathcal{B}_\sigma} \neg E_j\right] \leq e^{-\epsilon n(\alpha n^{-1/q})^i/6i} \leq \frac{1}{2}|\Xi|^{-|C|}$ since $\epsilon\alpha^i/6i \geq 4q(q+1)\log(|\Xi|)$.

If $\Delta \geq \eta$, then $\Pr\left[\bigwedge_{Q_j \in \mathcal{B}_\sigma} \neg E_j\right] \leq e^{-\eta^2/\Delta} \leq \exp\left(-|\mathcal{B}_\sigma|\alpha^i n^{-i/q} / \sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r}\right)$ by Equation 2 in Theorem 2.5. From Calculation 1 (see appendix), we know that $\alpha^i / \sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r}$ is at least α/i^2. Since, $|\mathcal{B}_\sigma| \geq \epsilon n/3i$, we get $\Pr\left[\bigwedge_{Q_j \in \mathcal{B}_\sigma} \neg E_j\right] \leq \exp(-\epsilon\alpha n^{1-i/q}/3i^3)$ which is at most $\frac{1}{2}|\Xi|^{-|C|}$.

Hence with probability at least $\frac{1}{2}|\Xi|^{-|C|}$, for every $\sigma \in \Xi^{|C|}$, there is a witness for $w_{\sigma,C}$ in the set sampled by the $\alpha n^{-1/q}$-sampler. Therefore, the set sampled by the $\alpha n^{-1/q}$-sampler is a super-witness against w with probability at least $\frac{1}{2}$, and by Observation 4.3 it is also a witness against w.

∎

We now prove the 1-sided test conversion result.

Proof of Theorem 4.1: Given a distribution μ which corresponds to a 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test for a property over Ξ^n, where $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$, we show for $\alpha = 15\ln|\Xi| \cdot q^4(q+1)^2/\epsilon$ that the $\alpha \cdot n^{-1/q}$-sampling distribution corresponds to a 1-sided $(\epsilon, 1/2)$-test for the same property.

We now refer to C_0 and $\mathcal{S}_0$ as defined in the statement of Lemma 3.2. First note that $|C_0| < 4q(q+1)\log(|\Xi|)n^{1-1/q} < \epsilon n/2$, where the last inequality follows from $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$. Therefore, by Lemma 2.12, $\mu(\mathcal{S}_0) \leq 1/(q+1)$, and so by Lemma 3.2 there exists an exact $(4q(q+1)\log(|\Xi|), i)$-constellation $(\mathcal{S}, C)$ for some $i \in [q]$ for which $\mu(\mathcal{S}) \geq 1/(q+1)$. Therefore, by Lemma 4.4, the theorem follows. ∎

V. Probabilistic formulas and test combinatorialization

Here we take a non-adaptive 2-sided test and make its structure more malleable to combinatorial arguments, with the main feature being that the new query distribution will be uniform over its support. At first, we define a

structure that can generally describe tests; we use this formulation to make the following arguments clearer and more succinct, which will also present them in their fullest possible generality.

Definition 5.1 (probabilistic constraints and formulas): A *probabilistic q-constraint* (over an alphabet Ξ) is a pair $C = (Q, S)$ where $Q \subseteq [n]$ is a *constraint set*, also called a *query set*, of size q, and S is a *satisfaction function* from $\Xi^{|Q|}$ to the real interval $[0, 1]$.

A *probabilistic q-formula* $P = (\mathcal{F}, \mu)$ is a set $\mathcal{F}$ of q-constraints, all with distinct constraint sets, along with a probability distribution μ over $\mathcal{F}$. We call it a (q, k)-formula if additionally $|\operatorname{Supp}(\mu)| \leq k$, in which case we can assume that $|\mathcal{F}| \leq k$.

When we drop the restriction on the sizes of the query sets of the constraints (even the restriction that they are all of the same size) then we just call P a *probabilistic formula*.

Given a word $w \in \Xi^n$ and a probabilistic formula P, the *satisfaction* of P by w is the average of the random variable that results from picking a constraint $(Q, S) \in \mathcal{F}$ according to μ and obtaining the value $S(w_Q)$. P is said to be *δ-sure for w* if its satisfaction by w is either at least $1 - \delta$ or at most δ.

The requirement for all sets corresponding to constraints being distinct allows us (given a particular formula $\mathcal{F}$) to identify the distribution μ with the corresponding distribution over subsets of $[n]$ only. This we will do throughout the sequel, but first let us justify this requirement.

Lemma 5.2: The requirement that the members of P have distinct query sets is without loss of generality.

Proof: If $C_1 = (Q, S_1)$ and $C_2 = (Q, S_2)$ are two constraints in a formula $P = (\mathcal{F}, \mu)$ (that for now does not satisfy the distinct set requirement), then we define $\mathcal{F}'$ by replacing them with $C = (Q, S)$ where $S = (\mu(C_1) \cdot S_1 + \mu(C_2) \cdot S_2)/(\mu(C_1) + \mu(C_2))$, and define the corresponding μ' by setting $\mu'(C) = \mu(C_1) + \mu(C_2)$. This preserves satisfaction values over all words w. We can continue doing this until there are no pairs left of constraints sharing the same query set. ∎

We shall henceforth abuse notation, and indeed refer to μ both as a distribution over $2^{[n]}$ and as a distribution over $\mathcal{F}$. Also, we shall make liberal use of the assumption (without loss of generality) that the support of μ is the entire $\mathcal{F}$ (otherwise we replace $\mathcal{F}$ with $\operatorname{Supp}(\mu) \subseteq \mathcal{F}$).

A non-adaptive 2-sided test or partial test can be described as follows.

Definition 5.3 (alternative definition of non-adaptive tests): Given the properties $L' \subseteq L \subseteq \Xi^n$, a *non-adaptive* 2-sided partial (ϵ, δ)-test for (L', L) is a probabilistic formula whose satisfaction over any $w \in L'$ is at least $1 - \delta$, while its satisfaction for any $w \in \Xi^n$ that is ϵ-far from L is at most δ.

If $L' = L$ then we just call it a 2-sided (ϵ, δ)-test for L.

If the test uses a q-formula then we may also call it an (ϵ, δ, q)-test, and if it uses a (q, k)-formula then we may call it an (ϵ, δ, q, k)-test.

To convert a non-adaptive test to this definition, we take μ to be the query distribution corresponding to the test, and set each pair (Q, S) so that S will describe the acceptance probability of the test given each possible outcome of its queries to Q.

We need the following technicality for the pairs (L', L) that we consider. It is safe to restrict our discussion to such pairs because otherwise there exists a trivial partial test.

Definition 5.4: Given two properties $L' \subseteq L \subseteq \Xi^n$, we say that the pair (L', L) is *ϵ-nontrivial* if there exist some word in L' and some word ϵ-far from L.

The purpose of this section is to show that all tests can be made to obey certain restrictions, at some reasonable cost for their parameters. To formulate the main lemma we need to define what these restrictions may be.

Definition 5.5 (restrictions on formulas and tests): A probabilistic formula P is said to be *zero-one* if all its constraints have the range $\{0, 1\}$ (instead of the whole interval).

P is said to be *β-equitable* if for every two constraints C_1 and C_2 in the support of the corresponding distribution μ, we have $\mu(C_1) \leq \beta\mu(C_2)$. In particular, for a 1-equitable formula the distribution μ is uniform over its support.

A q-formula P is said to be *combinatorial* if it is zero-one and equitable.

We use the same adjectives for tests. For example a test is called *combinatorial* if its corresponding formula is combinatorial.

We will prove the main combinatorialization lemma of this section following a sequence of steps. The easiest of these steps is making the corresponding formula zero-one.

Lemma 5.6: A formula P can be made into a zero-one formula P' without any change to its other parameters (including also its support size and equitability), so that for any input for which P was δ-sure about, P' will be 2δ-sure about and in the same direction.

Proof: For every constraint $C = (Q, S)$ in $\mathrm{Supp}(\mu)$, we replace it with $C' = (Q, S')$, where S' is defined so that $S'(v) = 0$ if $S(v) < \frac{1}{2}$, and otherwise $S'(v) = 1$. We leave μ "unmodified", that is, the new μ' is defined by having $\mu'(C') = \mu(C)$, in particular remaining identical as a distribution over query sets.

We present here the analysis for the case where the satisfaction of P by $w \in \Xi^n$ is at most δ. The case where it is at least $1 - \delta$ is symmetric. Given such a w, we set $\mathcal{F} = \mathrm{Supp}(\mu)$, and let $\mathcal{F}_2$ be the set of clauses whose satisfaction by w is at least $1/2$. Clearly $\mu(\mathcal{F}_2) \leq 2\delta$. The satisfaction of P' by w is now bounded by $0 \cdot \mu(\mathcal{F} \setminus \mathcal{F}_2) + 1 \cdot \mu(\mathcal{F}_2) \leq 2\delta$. ■

In the sequel we will need to analyze formulas conditioned on subsets of the original constraint set.

Definition 5.7: Given a probabilistic formula $P = (\mathcal{F}, \mu)$ and $\emptyset \neq \mathcal{F}' \subseteq \mathcal{F}$, the *$\mathcal{F}'$-conditioned* formula is $P' = (\mathcal{F}', \mu')$, where μ' is μ conditioned on the event that a member from $\mathcal{F}'$ was chosen.

The following fact about conditioned formulas is trivial.

Observation 5.8: Given $P = (\mathcal{F}, \mu)$, if $\mathcal{F}' \subseteq \mathcal{F}$ satisfies $\mu(\mathcal{F}') \geq \eta$, then for every input for which P was δ-sure about, the conditioned formula P' will be δ/η-sure about and in the same direction.

Proof: Again we analyze the case where the satisfaction of P by $w \in \Xi^n$ is at most δ, as the case where it is at least $1 - \delta$ is symmetric. For such w we write:

$$\sum_{(Q,S)\in\mathcal{F}'} \mu'(Q)S(w_Q) = \sum_{(Q,S)\in\mathcal{F}'} \mu(Q)S(w_Q)/\mu(\mathcal{F}') \leq \delta/\eta$$

where in the symmetric case we refer to $(1 - S(w_Q))$ instead of $S(w_Q)$. ■

We next prove a lemma (which like most lemmas of this section, holds also for formulas which are not tests), that allows us to move from β-equitable formulas all the way to 1-equitable ones. For making the transition cost not too high, we first prove a "quantization" step.

Lemma 5.9: A β-equitable formula $P = (\mathcal{F}, \mu)$ can be made into a formula $P' = (\mathcal{F}, \mu')$ for which μ' has at most $\log(2\beta)$ possible values, so that for any input for which P was δ-sure about, P' will be 2δ-sure about and in the same direction. Moreover, since P' has the same $\mathcal{F}$ and the same support, it preserves the original support size, query size, and zero-one property (if it existed) of P.

Proof: We first define $\tilde{\mu}$ by setting for every $C \in \mathcal{F}$ the value $\tilde{\mu}(C)$ to be 2^{-k_C}, where k_C is the largest integer for which $2^{-k_C} \geq \mu(C)$. Clearly for every C we have $\mu(C) \leq \tilde{\mu}(C) \leq 2\mu(C)$, and clearly $\tilde{\mu}$ has at most $\log(2\beta)$ possible values. However, it is not a probability measure, because it may be that $\tilde{\mu}(\mathcal{F}) > 1$. We thus set $\mu'(C) = \tilde{\mu}(C)/\tilde{\mu}(\mathcal{F})$ for every $C \in \mathcal{F}$.

Finally, if the satisfaction of P by w is at most δ, we write:

$$\sum_{(Q,S)\in\mathcal{F}} \mu'(Q)S(w_Q) \leq \sum_{(Q,S)\in\mathcal{F}} \tilde{\mu}(Q)S(w_Q) \leq 2 \cdot \sum_{(Q,S)\in\mathcal{F}} \mu(Q)S(w_Q) \leq 2\delta$$

where again the case of the satisfaction being at least $1 - \delta$ is symmetric. ■

Lemma 5.10: A β-equitable formula $P = (\mathcal{F}, \mu)$ can be made into a 1-equitable formula $P' = (\mathcal{F}', \mu')$, so that for any input for which P was δ-sure about, P' will be $2\delta\log(2\beta)$-sure about and in the same direction. Moreover, $\mathcal{F}' \subseteq \mathcal{F}$, so P' preserves the support size bound, query size bound, and possible zero-one property of P.

Proof: We first use Lemma 5.9 to move from P to $P'' = (\mathcal{F}, \mu'')$, where μ'' has at most $\log(2\beta)$ possible values, and any input for which P was δ-sure about, P'' is 2δ-sure about. Now there must be some $\eta \in (0, 1]$ so that $\mathcal{F}_\eta = \{C \in \mathcal{F} : \mu''(C) = \eta\}$ satisfies $\mu''(F_\eta) \geq 1/\log(2\beta)$.

We set $\mathcal{F}' = \mathcal{F}_\eta$ and make P' the formula of P'' conditioned on $\mathcal{F}'$. We finalize the proof by appealing to Observation 5.8. ∎

Specifically for (partial) tests, we next show some correlation between subsets "covering" few indexes and probability. The following lemma will also be used when constructing sets of pompoms to prove the main conversion result from 2-sided tests to sampling tests.

Lemma 5.11: If a formula $P = (\mathcal{F}, \mu)$ corresponds to an $(\epsilon/2, \delta)$-test for (L', L) which is ϵ-nontrivial, and $\mathcal{F}' \subseteq \mathcal{F}$ is such that the union of its corresponding query sets occupies at most $\epsilon n/2$ indexes from $[n]$, then $\mu(\mathcal{F}') \leq 2\delta$.

Proof: Let $T = \bigcup_{(Q,S)\in\mathcal{F}'} Q$, $u \in \Xi^n$ be a word in L', $w \in \Xi^n$ be ϵ-far from L, and $v = w_{u_T,T}$ be such that $v_T = u_T$ and $v_{[n]\setminus T} = w_{[n]\setminus T}$. By the triangle inequality, v is $\epsilon/2$-far from L, and so

$$\sum_{(Q,S)\in\mathcal{F}'} \mu(Q)S(u_Q) = \sum_{(Q,S)\in\mathcal{F}'} \mu(Q)S(v_Q) \leq \delta$$

since this bounds the satisfaction of P by v.

On the other hand, the satisfaction of P by u is at least $1-\delta$, and so we obtain

$$1-\delta \leq \sum_{(Q,S)\in\mathcal{F}'} \mu(Q)S(u_Q) + \sum_{(Q,S)\in\mathcal{F}\setminus\mathcal{F}'} \mu(Q)S(u_Q) \leq \delta + (1-\mu(\mathcal{F}'))$$

which necessarily means that $\mu(\mathcal{F}') \leq 2\delta$. ∎

Using the above lemma we can show that tests with a small enough support size can be made into equitable ones.

Lemma 5.12: For $\delta < \frac{1}{8}$, an $(\epsilon/2, \delta, q, \alpha n)$-test for (L', L) which is ϵ-nontrivial can be made into a β-equitable $(\epsilon/2, 2\delta, q, \alpha n)$-test for (L', L), for $\beta = 8q\alpha/\epsilon$. This transformation also preserves the zero-one property if it existed.

Proof: Let $P = (\mathcal{F}, \mu)$ be the formula corresponding to the test. First we let $\mathcal{F}_0 = \{C \in \mathcal{F} : \mu(C) \leq 1/4\alpha n\}$. Clearly $\mu(\mathcal{F}_0) \leq \frac{1}{4}$. Now let $\mathcal{F}_1 = \{C \in \mathcal{F} : \mu(C) \geq 2q/\epsilon n\}$. Clearly $|\mathcal{F}_1| \leq \epsilon n/2q$. Hence $|\bigcup_{(Q,S)\in\mathcal{F}_1} Q| \leq \epsilon n/2$, and so by Lemma 5.11 we have $\mu(\mathcal{F}_1) \leq 2\delta \leq \frac{1}{4}$.

Setting $\mathcal{F}' = \mathcal{F} \setminus (\mathcal{F}_0 \cup \mathcal{F}_1)$, we get $\mu(\mathcal{F}') \geq \frac{1}{2}$. Setting P' to be the conditioning of P to $\mathcal{F}'$, we obtain by Observation 5.8 that it is an $(\epsilon/2, 2\delta, q, \alpha n)$-test for (L', L). Moreover, since for every $C \in \mathcal{F}'$ we have $1/4\alpha n < \mu(C) < 2q/\epsilon n$, we get that the resulting test is β-equitable for $\beta = 8q\alpha/\epsilon$. ∎

We now have nearly all the ingredients we need. The final one is a way to convert a general test to one whose support size is linear in n, which the following lemma provides even for formulas that are not necessarily tests.

Lemma 5.13: Any q-formula $P = (\mathcal{F}, \mu)$ can be made into a $(q, \alpha n)$-formula P' for $\alpha = \delta^{-2}\log(|\Xi|)$, with the condition that any $w \in \Xi^n$, for which P was δ-sure about, P' will be 2δ-sure about and in the same direction. This also preserves the zero-one property if it exists.

Proof: To produce the new formula, we take $r = \delta^{-2}\log(|\Xi|) \cdot n$ samples $(Q_1, S_1), \ldots, (Q_r, S_r)$ from $\mathcal{F}$ by independently drawing each sample according to μ. For $w \in \Xi^n$ we set $\eta_w = (\sum_{i=1}^r S_i(w_{Q_i}))/r$. We also let $\eta = \sum_{(Q,S)\in\mathcal{F}} \mu(Q)S(w_Q)$ denote the satisfaction of P by w. Let Y_i denote the random variable $S_i(w_{Q_i})$, and set $Y = \sum_{i=1}^r Y_i/r$. Note that $\mathrm{E}[Y_i] = \sum_{(Q,S)\in\mathcal{F}} \mu(Q)S(w_Q) = \eta$. Thus also $\mathrm{E}[Y] = \eta$, and by Lemma 2.2 we have that the probability for $|\eta_w - \eta| > \delta$ is bounded by $2e^{-2r\delta^2} \leq \frac{1}{2}|\Xi|^{-n}$.

Thus, with probability at least $\frac{1}{2}$, the obtained sequence is such that for all $w \in \Xi^n$ we have that the difference between η_w and η is at most δ. We fix such a sequence $(Q_1, S_1), \ldots, (Q_r, S_r)$. To define $P' = (\mathcal{F}', \mu')$, we set $\mathcal{F}'$ to be the set of clauses appearing in $(Q_1, S_1), \ldots, (Q_r, S_r)$, where for $C \in \mathcal{F}'$ we set $\mu'(C)$ to be the number of times it appeared in the sequence, divided by r. ∎

Now we are finally ready to prove the main combinatorialization result of this section.

Lemma 5.14 (combinatorialization lemma): Any (partial) $(\epsilon/2, \delta, q)$-test for (L', L) which is ϵ-nontrivial can be made into a combinatorial $(\epsilon/2, \delta', q, \alpha n)$-test, where $\alpha = \delta^{-2}\log(|\Xi|)$ and $\delta' = 16\delta\log(16q\delta^{-2}\log(|\Xi|)/\epsilon)$.

Proof: Setting P to be the formula corresponding to the test, we can assume that $\delta < \frac{1}{16}$, as otherwise we can just ignore P and provide a "test" that is satisfied by all inputs. We perform the following sequence of steps.

- Use Lemma 5.13 to make it into a formula P_1 corresponding to an $(\epsilon/2, 2\delta, q, \alpha n)$-test for $\alpha = \delta^{-2}\log(|\Xi|)$.
- Use Lemma 5.12 to make P_1 into a formula P_2 that is an $(8q\delta^{-2}\log(|\Xi|)/\epsilon)$-equitable $(\epsilon/2, 4\delta, q, \alpha n)$-test. This is the only step that requires the formula to correspond to a test.
- Use Lemma 5.10 to make P_2 into P_3 which is a 1-equitable $(\epsilon/2, 8\delta\log(16q\delta^{-2}\log(|\Xi|)/\epsilon), q, \alpha n)$-test.
- Finally use Lemma 5.6 to make P_3 a combinatorial partial $(\epsilon/2, 16\delta\log(16q\delta^{-2}\log(|\Xi|)/\epsilon), q, \alpha n)$-test for (L', L), whose formula we denote by P'.

The final formula P' is the required test. ∎

Before concluding this section, we note that by analyzing the only place in the proof where we used that P is a test, we can formulate the following more general combinatorialization lemma, that can be of independent use.

Lemma 5.15 (general combinatorialization): If P is a probabilistic q-formula resolving with δ confidence a promise problem, for which there are both "yes" instances, and words where every $\epsilon/2$-close word is a "no" instance, then P can be made into a combinatorial $(q, \alpha n)$ formula resolving the same promise problem with δ' confidence, where α and δ' are as in Lemma 5.14.

To conclude this section, we combine Lemma 5.14 with an amplification technique to show how general 2-sided $(\epsilon/2, \delta, q)$-tests can be converted to combinatorial tests.

Lemma 5.16: A partial $(\epsilon/2, \delta, q)$-test for (L', L) which is ϵ-nontrivial can be converted to a combinatorial partial $(\epsilon/2, 1/10(q'+1), q', (q'+1)^2\log(|\Xi|)(\log((q'+1)\log(|\Xi|)/\epsilon))^2 n)$-test for (L', L), where $q' = O(q\log(q)\log\log(\log(|\Xi|)/\epsilon)\log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2)$.

Proof: We first perform 2-sided amplification: We use $100\log(q)\log\log(\log(|\Xi|)/\epsilon)\log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2$ repetitions of the original test, taking the majority vote to obtain an $(\epsilon/2, 1/10^{12}(\tilde{q}+1)\log((\tilde{q}+1)\log(|\Xi|)/\epsilon), \tilde{q})$-test for $\tilde{q} = O(q\log(q)\log\log(\log(|\Xi|)/\epsilon)\log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2)$. Now we use Lemma 5.14 on the amplified test, and obtain a combinatorial $(\epsilon/2, 1/10^8(\tilde{q}+1), \tilde{q}, 10^{12}(\tilde{q}+1)^2\log(|\Xi|)(\log((\tilde{q}+1)\log(|\Xi|)/\epsilon))^2 n)$-test. To obtain the lemma's conclusion, we artificially increase the number of queries from $\tilde{q}$ to $q' = 10^7\tilde{q}$. ∎

We note that the dependency of q' above on $\log\log(\log(|\Xi|)/\epsilon)$ implies that a constant power of n is guaranteed only for properties for which the alphabet does not depend on n. However, it is unlikely to destroy sublinearity by itself even for a variable alphabet setting, because already for $|\Xi| = 2^n$ there are examples with no sublinear sampling tests at all by [4]. Our main theorem will cease to work, due to its minimum n requirement, when $|\Xi|$ is larger than an exponential in some power of n (that is linear in $1/q$).

VI. A CONVERSION OF A 2-SIDED TEST TO A 2-SIDED SAMPLING TEST

Here we prove that if the properties $L' \subseteq L \subseteq \Xi^n$ admit a 2-sided test with a constant number of queries for (L', L), then there is a corresponding 2-sided p-sampling test where p corresponds to a constant negative power of n. Specifically we prove the following.

Theorem 6.1: Let $\alpha = 10^3\ln(|\Xi|)(q+1)^2/\epsilon$ for any $q \geq 3$, and $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$. if (L', L) is ϵ-nontrivial and admits a 2-sided combinatorial partial $(\epsilon/2, 1/10(q+1), q, (q+1)^2\log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n)$-test, then it also admits a p-sampling 2-sided $(\epsilon, 1/10)$-test such that $p = \alpha n^{-1/q^2}$.

As with the proof of the 1-sided case, we set μ to be a distribution of the test, and find for it pompoms that cover every possible assignment to a common core set C. Here however there are many pompoms involved, and they serve all assignments to C at once, because we need to cover enough of the "weight" of the distribution μ. Also, the pompoms are not necessarily of witnesses, but rather of query sets; we will use them to approximate for every assignment to C the "amount" of query sets that would cause its rejection (hence they need to cover sufficient weight). We need the test to be combinatorial, i.e., that μ is uniform over its support, for exactly one proof step: The sampling test will approximate the *number* of rejecting query sets, and only for a uniform μ will this correspond to the rejection *probability* of the original test.

Let us formally define the set of pompoms that we will use.

Definition 6.2: Given a distribution μ over sets of size q, a set J of i-pompoms made from members of $\mathrm{Supp}(\mu)$ is *discerning* for μ if the following holds:

1) $\mu(\mathcal{I}) \geq \frac{1}{2(q+1)}$, where $\mathcal{I} = \bigcup_{\mathcal{W} \in J} \mathcal{W}$ is the union of all the i-pompoms in J.
2) Every i-pompom in J has cardinality exactly $\epsilon \cdot n^{1-(i-1)/q}/3i$.
3) There exists $C \subseteq [n]$ of size at most $q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-i/q}$ that is a core of all the i-pompoms in J.

Next we define and show how a pompom of such a set can be used in a proof of something like Theorem 6.1.

Definition 6.3: Given a q-formula $P = (\mathcal{F}, \mu)$ over Ξ^n, an i-pompom $\mathcal{W} \subseteq \mathrm{Supp}(\mu)$ of size $\epsilon \cdot n^{1-(i-1)/q}/3i$ with core C of size at most $q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-i/q}$, a word $w \in \Xi^n$, a possible assignment $\sigma \in \Xi^{|C|}$ to C, and a query set $U \subseteq [n]$, the *approximated satisfiability of $\mathcal{W}$ by σ with respect to w* is defined to be the value $\gamma_{\sigma,U,\mathcal{W}}$ obtained in the following manner.

Set $\mathcal{W}_U = \{Q \in \mathcal{W} : Q \setminus C \subseteq U\}$ (i.e., take the set of members of $\mathcal{W}$ whose indexes outside C are contained in U), and then take the average $\gamma_{\sigma,U,\mathcal{W}} = (\sum_{\{(Q,S) \in \mathcal{F} : Q \in \mathcal{W}_U\}} S((w_{\sigma,C})_Q))/|\mathcal{W}_U|$, which we arbitrarily set to $\frac{1}{2}$ if $\mathcal{W}_U = \emptyset$.

An explanation to the above definition: Assume that U is a set of queries that we have made. We would like to assess the assignment σ to C, with respect to what U tells us about w outside of C. Given the i-pompom $\mathcal{W}$, we want to approximate the relative weight of the members of $\mathcal{W}$ for which the corresponding constraints accept $w_{\sigma,C}$. We do so by restricting ourselves to the members of $\mathcal{W}_U$, for which we can tell by querying U whether they accept $w_{\sigma,C}$ or not. We ignore all aspects of μ apart from its support, because we will assume that it is uniform over $\mathrm{Supp}(\mu)$ (i.e., that the formula P corresponds to a combinatorial test). This assumption is essential to show that a set U chosen according to a sampling distribution will indeed yield with high probability a good approximation.

Note that $\gamma_{\sigma,[n],\mathcal{W}}$ is the true acceptance average of the pompom $\mathcal{W}$. We now prove that the sampling distribution with high probability provides a U such that $\gamma_{\sigma,U,\mathcal{W}}$ approximates $\gamma_{\sigma,[n],\mathcal{W}}$.

Lemma 6.4: Let $q \geq 3$, $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, $\alpha = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$ and $w \in \Xi^n$. Suppose that the formula $P = (\mathcal{F}, \mu)$, the i-pompom $\mathcal{W}$ and its core C, and the words w and σ are as per the requirements of Definition 6.3, and additionally that P is combinatorial. Then with probability at least $1 - \frac{1}{100}|\Xi|^{-|C|}$, a set U drawn according to the $\alpha \cdot n^{-1/q^2}$-sampling distribution satisfies $|\gamma_{\sigma,U,\mathcal{W}} - \gamma_{\sigma,[n],\mathcal{W}}| \leq \frac{1}{10}$.

Proof: Let us first analyze which members of $\mathcal{W}$ get into $\mathcal{W}_U$. Since $\{Q \setminus C : Q \in \mathcal{W}\}$ is a family of disjoint sets of size i, the choice of U means that every $Q \in \mathcal{W}$ becomes a member of $\mathcal{W}_U$ with probability exactly $\alpha^i \cdot n^{-i/q^2}$, independently of other members of $\mathcal{W}$. We now refer to Lemma 2.4 (where $(\gamma_1, \ldots, \gamma_m)$ there are the satisfaction values $S((w_{\sigma,C})_Q)$ for $(Q,S) \in \mathcal{F}$ such that $Q \in \mathcal{W}$), which bounds the probability for $|\gamma_{\sigma,U,\mathcal{W}} - \gamma_{\sigma,[n],\mathcal{W}}| > \frac{1}{10}$ by $e^{-10^{-3}\alpha^i \cdot n^{-i/q^2} \cdot \epsilon \cdot n^{1-(i-1)/q}/3i}$. Calculation 2 bounds this by $\frac{1}{100}|\Xi|^{-q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-i/q}} \leq \frac{1}{100}|\Xi|^{-|C|}$. ∎

From the above lemma we formulate a way of approximating all pompoms in a discerning set J, assuming that we have knowledge of J, the common core set C, and of course the original combinatorial test $(\mathcal{F}, \mu)$.

Lemma 6.5: Assume that $q \geq 3$, $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$ and $w \in \Xi^n$. Let $P = (\mathcal{F}, \mu)$ be a combinatorial $(\epsilon/2, 1/10(q+1), q, (q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n)$-test for (L', L) which is ϵ-nontrivial, let J be a discerning set of i-pompoms for it with core C, and let U be chosen by the $\alpha \cdot n^{-1/q^2}$-sampling distribution where $\alpha = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$. With probability at least $\frac{9}{10}$, for every $\sigma \in \Xi^{|C|}$ it holds that $|\frac{1}{|J|}\sum_{\mathcal{W} \in J} \gamma_{\sigma,U,\mathcal{W}} - \frac{1}{|J|}\sum_{\mathcal{W} \in J} \gamma_{\sigma,[n],\mathcal{W}}| \leq \frac{1}{5}$.

Proof: For a fixed $\sigma \in \Xi^{|C|}$, by using Lemma 6.4 and Markov's inequality, we obtain that with probability at most $\frac{1}{10}|\Xi|^{-|C|}$ we have more than $\frac{1}{10}|J|$ instances $\mathcal{W} \in J$ for which $|\gamma_{\sigma,U,\mathcal{W}} - \gamma_{\sigma,[n],\mathcal{W}}| > \frac{1}{10}$. Therefore, with probability at least $1 - \frac{1}{10}|\Xi|^{-|C|}$ (noting that every γ value is always between 0 and 1) the following holds:

$$|\frac{1}{|J|}\sum_{\mathcal{W} \in J} \gamma_{\sigma,U,\mathcal{W}} - \frac{1}{|J|}\sum_{\mathcal{W} \in J} \gamma_{\sigma,[n],\mathcal{W}}| \leq \frac{1}{|J|}\sum_{\mathcal{W} \in J} |\gamma_{\sigma,U,\mathcal{W}} - \gamma_{\sigma,[n],\mathcal{W}}| \leq \frac{1}{10} + \frac{1}{10} = \frac{1}{5}$$

A union bound over the bad events for every possible $\sigma \in \Xi^{|C|}$ concludes the proof. ■

We now show that a constellation as defined in Definition 3.1 that encompasses at least $1/(q+1)$ of the query sets implies a discerning set of pompoms. Later we will use Lemma 3.2 to find the required constellation, just as we did for the case of 1-sided tests.

Lemma 6.6: Let $i \in [q]$ and $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, (L', L) be ϵ-nontrivial, $P = (\mathcal{F}, \mu)$ be a combinatorial $(\epsilon/2, 1/10(q+1), q, (q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n)$-test for (L', L), and let $(\mathcal{S}, C)$ be a $(q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2, i)$-constellation for μ. Then there exists a set J of i-pompoms that is discerning for P with core C.

Proof: We extract pompoms from $\mathcal{S}$ one by one. Denoting by $\mathcal{T}$ the family of sets remaining from $\mathcal{S}$ after already extracting some pompoms, we claim that as long as $\mu(\mathcal{T}) > \frac{1}{2(q+1)}$, we can extract another i-pompom $\mathcal{W}$ of size $\epsilon \cdot n^{1-(i-1)/q}/3i$ with core C from $\mathcal{S}$, which we then subtract from $\mathcal{T}$ and make into a new member of J. Assuming the claim holds, the process stops only when J becomes such that Item 1 of Definition 6.2 holds, because we started with a set $\mathcal{S}$ of weight at least $\frac{1}{q+1}$. Also, Item 3 of Definition 6.2 follows from Condition 1 of Definition 3.1 (regarding $\mathcal{S}$ and C), while Item 2 follows from the procedure described above.

It thus remains to show that given a $(q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2, i)$-constellation $(\mathcal{T}, C)$ for which $\mu(\mathcal{T}) > \frac{1}{2(q+1)}$, an i-pompom $\mathcal{W} \subseteq \mathcal{T}$of size $\epsilon \cdot n^{1-(i-1)/q}/3i$ with core C exists. Since $\mu(\mathcal{T}) > \frac{1}{2(q+1)}$, by Lemma 5.11, we have that $|\bigcup_{Q \in \mathcal{T}} Q| \geq \epsilon n/2$. On the other hand we observe that $|C| \geq \epsilon n/6$ since $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$. Thus we can use Observation 3.4 (with $\beta = \epsilon \cdot n/3$) to obtain the required i-pompom $\mathcal{W}$ of cardinality $\epsilon \cdot n^{1-(i-1)/q}/3i$ (if its cardinality is larger than that, then we arbitrarily reduce it). ■

Now we prove the main result of this section.

Proof of Theorem 6.1: Given a $(\epsilon/2, 1/10(q+1), q, (q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n)$-test for (L', L) that is combinatorial, where $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, we construct for $\alpha = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$ a 2-sided $(\epsilon, \frac{1}{10})$-test for (L', L) that uses the $\alpha \cdot n^{-1/q^2}$-sampling distribution.

We will use for the distribution μ of the combinatorial test (uniform over its family of possible query sets) Lemma 3.2. Referring to C_0 and $\mathcal{S}_0$ as defined in the statement of this lemma, $|C_0| < q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-1/q} < \epsilon n/2$, where the last inequality follows from $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$. Therefore, by Lemma 5.11, $\mu(\mathcal{S}_0) \leq 1/(q+1)$, so by Lemma 3.2 there exists a $(q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2, i)$-constellation $(\mathcal{S}, C)$ for some $i \in Q$ satisfying $\mu(\mathcal{S}) \geq 1/(q+1)$.

Moreover, we note that such a constellation can indeed be computed from only the knowledge of $\mathrm{Supp}(\mu)$. We then use Lemma 6.6 (which is also constructive) to obtain the discerning set J of i-pompoms.

The test proceeds as follows. Given the set U produced by the $\alpha \cdot n^{-1/q^2}$-sampling distribution, we query all of it. Then, for every $\sigma \in \Xi^{|C|}$ and every $\mathcal{W} \in J$, we calculate $\gamma_{\sigma,U,\mathcal{W}}$ using our queries, and then calculate $\gamma_{\sigma,U} = \frac{1}{|J|} \sum_{\mathcal{W} \in J} \gamma_{\sigma,U,\mathcal{W}}$ for every σ. If there was a $\sigma \in \Xi^{|C|}$ for which $\gamma_{\sigma,U} > \frac{1}{2}$, then we accept the input, and otherwise we reject it.

It remains to prove that this is indeed a correct test for (L', L). Set $\mathcal{I} = \bigcup_{\mathcal{W} \in J} \mathcal{W}$ as per Definition 6.2. Since $\mu(\mathcal{I}) \geq \frac{1}{2(q+1)}$, if u is any word for which the original test was $\frac{1}{10(q+1)}$-sure about, then the conditioning of the test to the set of constraints corresponding to the members of $\mathcal{I}$ will be $\frac{1}{5}$-sure for u by Observation 5.8. Now, since μ is uniform over its support, for any $\sigma \in \Xi^{|C|}$, the satisfaction of the original test conditioned on $\mathcal{I}$ by $w_{\sigma,C}$ is identical to the average $\gamma_{\sigma,[n]} = \frac{1}{|J|} \sum_{\mathcal{W} \in J} \gamma_{\sigma,[n],\mathcal{W}}$. In turn, Lemma 6.5 guarantees that with probability at least $\frac{9}{10}$, for all such σ we have $|\gamma_{\sigma,U} - \gamma_{\sigma,[n]}| \leq \frac{1}{5}$. Assume from now on that this event has indeed occurred.

If w was a word in L', then the original test accepted it with probability at least $1 - \frac{1}{10(q+1)}$, and hence for $\sigma = w_C$ (for which $w_{\sigma,C} = w$) we have $\gamma_{\sigma,[n]} \geq \frac{4}{5}$ and hence $\gamma_{\sigma,U} \geq \frac{3}{5} > \frac{1}{2}$, and the sampling test will accept on account of this σ.

On the other hand, if w was a word ϵ-far from L, then for every $\sigma \in \Xi^{|C|}$, the word $w_{\sigma,C}$ is $\epsilon/2$-far from L (recall that in particular $|C| < \epsilon n/2$), and so the original test will accept it with probability at most $\frac{1}{10(q+1)}$. Hence $\gamma_{\sigma,[n]} \leq \frac{1}{5}$ for every such σ, and hence $\gamma_{\sigma,U} \leq \frac{2}{5} < \frac{1}{2}$. This means that the sampling test will reject, as

there will be no σ on whose account the test can accept. ∎

VII. Implications of our results

The following corollaries result respectively from Theorem 4.1 and Theorem 6.1, considering that a multitest scheme (as described in the introduction) immediately leads to a test for a union of the properties.

Corollary 7.1: Let $q = 60q' \log(q')/(1-\delta)$ and $\alpha_1 = 15 \ln|\Xi| \cdot q^4(q+1)^2/\epsilon$. For every $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$, if $L \subseteq \Xi^n$ is the union of r properties $L_1, \ldots, L_r$, each having a 1-sided $(\epsilon/2, \delta, q')$-test where $r \leq 2^{(\alpha_1)^{-1} n^{(q^{-1}-\gamma)}-1}$, then L has a non-adaptive 1-sided $(\epsilon, 1/2)$-test with query complexity $O(n^{1-\gamma})$.

Proof: First, we use Lemma 2.11 to convert the $(\epsilon/2, \delta, q')$-test for every L_i to a non-adaptive 1-sided $(\epsilon/2, 1/2(q+1), q, 4(q+1)\log(|\Xi|)n)$-test where $q = 60q'\log(q')/(1-\delta)$. We then use Theorem 4.1 to obtain an $\alpha_1 n^{-1/q}$-sampling 1-sided $(\epsilon/2, 1/2)$-test for every L_i. We then amplify the probability, by repeating the test and rejecting if any of the runs rejected, to obtain a $\log(2r)\alpha_1 n^{-1/q}$-sampling 1-sided $(\epsilon/2, 1/2r)$-test for every L_i. We construct a multitest for $L_1, \ldots, L_r$, which reuses the same queries for each sample-based test, and from it derive the test for $\bigcup_{i=1}^r L_i$. Since r is at most $2^{(\alpha_1)^{-1} n^{(q^{-1}-\gamma)}-1}$, this gives a non-adaptive 1-sided $(\epsilon, 1/2)$-test with query complexity $O(n^{1-\gamma})$. ∎

Corollary 7.2: Let $q = 10^9 q' \log(q') \log\log(\log(|\Xi|/\epsilon)) \log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2$ and let $\alpha_2 = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$. For every $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, if a property $L \subseteq \Xi^n$ is the union of r properties $L_1, \ldots, L_r$, each having a 2-sided $(\epsilon/2, \delta, q')$-test where $r \leq 2^{(10\alpha_2)^{-1} n^{(q^{-2}-\gamma)}}$, then L has a non-adaptive 2-sided $(\epsilon, 1/10)$-test with query complexity $O(n^{1-\gamma})$.

Proof: Use Lemma 5.16 to convert the $(\epsilon/2, \delta, q')$-test for each L_i to a combinatorial 2-sided $(\epsilon/2, 1/10(q+1), q, (q+1)^2\log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n)$-test for L_i, for $q = 10^9 q' \log(q')\log\log(\log(|\Xi|)/\epsilon)\log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2)$. Then we use Theorem 6.1 to convert each of these tests to an $\alpha_2 n^{-1/q^2}$-sampling 2-sided $(\epsilon, 1/10)$ test. Now, we convert them to $10\log(r)\alpha_2 n^{-1/q^2}$-sampling 2-sided $(\epsilon, 1/10r)$-tests by repeating each test $10\log r$ times independently and taking the majority vote. We construct a multitest for $L_1, \ldots, L_r$, which reuses the same queries for each sample-based test, and from it derive the test for $\bigcup_{i=1}^r L_i$. Since r is at most $2^{(10\alpha_2)^{-1} n^{(q^{-2}-\gamma)}}$, this gives the 2-sided $(\epsilon, 1/10)$-test with query complexity $O(n^{1-\gamma})$. ∎

Definition 7.3 (following Definition 2.1 of [8]): A *Merlin-Arthur proof of proximity* ($\mathcal{MAP}$) for a property $L \subseteq \Xi^n$, with proximity parameter ϵ, query complexity q and proof complexity p, consists of a probabilistic algorithm V, called the verifier, that is given a proof string $\pi \in \Xi^p$; in addition, it is given oracle access to a word $w \in \Xi^n$, to which it is allowed to make up to q queries. The verifier satisfies the following two conditions:

1) *Completeness*: For every $w \in L$, there exists a string $\pi \in \Xi^p$ (referred to as a *proof* or *witness*) such that $\Pr[V(n, \epsilon, w, \pi) = 1] \geq 2/3$.
2) *Soundness*: For every $w \in \Xi^n$ which is ϵ-far from L, and any $\pi \in \Xi^p$, $\Pr[V(n, \epsilon, w, \pi) = 1] \leq 1/3$.

If the completeness condition holds with probability 1, then we say that the $\mathcal{MAP}$ has 1-sided error, and otherwise we say that it has 2-sided error. Also, we may say that it is *non-adaptive* if it makes its queries to w based only on π, before receiving any responses from w.

For our purposes, we note that the proof of a $\mathcal{MAP}$ scheme for a property L induces a decomposition of L into sets whose union is L, each admitting a corresponding partial testing algorithm. Specifically, for every $w \in L$ we define Π_w to be any non-empty subset of the set of proofs $\pi \in \Xi^p$ that make the verifier accept w with the required probability. Then, for every $\pi \in \Xi^p$ we set $L_\pi = \{w \in L : \pi \in \Pi_w\}$ (it may be the case that some L_π are empty).

Under this interpretation, for a word in the property, the proof π is simply an indicator that the word belongs to L_π. Thus, the verifier of the $\mathcal{MAP}$ scheme can be seen as receiving as input a proof π and then running a partial test for (L_π, L). Consequently, the existence of a $\mathcal{MAP}$ scheme with query complexity q and proof complexity p for a property L is the same as having a family of $|\Xi|^p$ properties $\{L_\pi : \pi \in \Xi^p\}$ such that $L = \bigcup_{\pi \in \Xi^p} L_\pi$, and there exists a partial test for every pair (L_π, L).

Similarly to Corollary 7.2, only using the validity of Theorem 6.1 for partial tests as well, we obtain:

Corollary 7.4: Let $q = 10^9 q' \log(q') \log\log(\log(|\Xi|/\epsilon)) \log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2$ and let $\alpha_2 = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$. For every $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, if a property $L \subseteq \Xi^n$ has no non-adaptive 2-sided $(\epsilon/2, 1/10)$-tests with query complexity $o(n^{1-\gamma})$, then every 2-sided $\mathcal{MAP}$ scheme for L, that has query complexity q', has proof complexity $\Omega(n^{q^{-2}-\gamma}/10\alpha_2)$.

Although Theorem 4.1 was stated and proved for (non-partial) 1-sided tests only, it can also be made to work for partial tests, and to give a corollary with an improved bound for this case.

Corollary 7.5: Let $q = 60q' \log(q')/(1-\delta)$ and let $\alpha_1 = 15 \ln|\Xi| \cdot q^4(q+1)^2/\epsilon$. For every $n > (24q(q+1)(\log(|\Xi|))^2/\epsilon)^q$, if a property $L \subseteq \Xi^n$ has no non-adaptive 1-sided $(\epsilon/2, 1/2)$-tests with query complexity $o(n^{1-\gamma})$, then every 1-sided $\mathcal{MAP}$ scheme for L, that has query complexity q', has proof complexity $\Omega(n^{q^{-1}-\gamma}/\alpha_1 - 1)$.

We note some concrete applications of the above results.

- In [12], it was shown that there exists a language L with logarithmic space complexity that satisfies the following: every non-adaptive 2-sided $(\epsilon/2, \delta)$-test for $L \cap \{0,1\}^n$ has query complexity $\Omega(n)$. By Corollary 7.2, this means that for every fixed $\gamma > 0$, $q = 10^9 q' \log(q') \log\log(\log(2/\epsilon)) \log(1/(\frac{1}{2}-\delta))/(\frac{1}{2}-\delta)^2$, α_2 as defined above (with $|\Xi| = 2$), and large enough n, $L \cap \{0,1\}^n$ cannot be the union of less than $2^{(10\alpha_2)^{-1} n^{(q^{-2}-\gamma)}}$ properties over $\{0,1\}^n$ each having a 2-sided $(\epsilon/2, \delta, q')$-test. By Corollary 7.4, L does not have a 2-sided $\mathcal{MAP}$ with query complexity q' and proof complexity $o(n^{q^{-2}-\gamma}/10\alpha_2)$. Similarly, such conclusions apply to the properties of the small CNF formula that was studied in [13].
- Our result also applies to the sparse graph property of 3-colorability, which in [14] is shown to have a linear 2-sided test query complexity. Note that in the sparse graph model the size of the alphabet Ξ is n, but this is still small enough for our results to provide non-trivial conclusions against decomposability.
- According to the results in [15] and [16] respectively, every property defined by a constant width read-once branching program or a constant arity read-once Boolean formula is testable. Hence our results imply that properties whose testing requires $\Omega(n^{1-\gamma})$ many queries, for a corresponding γ, cannot be written as the union of a small number of properties that have such representations.

REFERENCES

[1] O. Goldreich and L. Trevisan, "Three theorems regarding testing graph properties," *Random Structures & Algorithms*, vol. 23, no. 1, pp. 23–57, 2003.

[2] O. Goldreich, S. Goldwasser, and D. Ron, "Property testing and its connection to learning and approximation," *J. ACM*, vol. 45, pp. 653–750, July 1998.

[3] A. Bhattacharyya, E. Grigorescu, and A. Shapira, "A unified framework for testing linear-invariant properties," *Random Structures & Algorithms*, vol. 46, no. 2, pp. 232–260, 2015.

[4] O. Goldreich and D. Ron, "On sample-based testers," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 337–345.

[5] ——, "On proximity-oblivious testing," *SIAM J. Comput*, vol. 40, no. 2, pp. 534–566, 2011. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/100789646

[6] E. Fischer, Y. Goldhirsh, and O. Lachish, "Partial tests, universal tests and decomposability," in *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM, 2014, pp. 483–500.

[7] N. Alon, "Testing subgraphs in large graphs," *Random Structures & Algorithms*, vol. 21, no. 3-4, pp. 359–370, 2002.

[8] T. Gur and R. D. Rothblum, "Non-interactive proofs of proximity," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ser. ITCS '15. ACM, 2015, pp. 133–142.

[9] P. Erdős and R. Rado, "Intersection theorems for systems of sets," *J. London Math. Soc.*, vol. 35, pp. 85–90, 1960.

[10] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[11] S. Janson, T. Łuczak, and A. Ruciński, "An exponential bound for the probability of nonexistence of a specified subgraph in a random graph," in *Random Graphs*, M. Karonski *et al.*, Eds. Wiley, 1990, pp. 73–88.

[12] O. Lachish, I. Newman, and A. Shapira, "Space complexity vs. query complexity," *Computational Complexity*, vol. 17, no. 1, pp. 70–93, 2008.

[13] E. Ben-Sasson, P. Harsha, and S. Raskhodnikova, "Some 3CNF properties are hard to test," *SIAM Journal on Computing*, vol. 35, no. 1, pp. 1–21, 2005.

[14] A. Bogdanov, K. Obata, and L. Trevisan, "A lower bound for testing 3-colorability in bounded-degree graphs," in *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*. IEEE, 2002, pp. 93–102.

[15] I. Newman, "Testing membership in languages that have small width branching programs," *SIAM Journal on Computing*, vol. 31, no. 5, pp. 1557–1570, 2002.

[16] E. Fischer, Y. Goldhirsh, and O. Lachish, "Testing formula satisfaction," in *Algorithm Theory – SWAT 2012*, ser. Lecture Notes in Computer Science. Springer, 2012, vol. 7357, pp. 376–387.

APPENDIX

This appendix is for calculations that are too long and bothersome to be put where they are originally used.

Calculation 1: For a positive integer q and any $0 < \epsilon \leq 1$, $i \in [q]$, and $\alpha = 15 \ln |\Xi| \cdot q^4 (q+1)^2/\epsilon$, we write:

$$\frac{\alpha^i}{\sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r}} = \frac{\alpha^i}{(1+\alpha)^i - \alpha^i - 1}.$$

For all $r < i$,

$$\binom{i}{r} \alpha^{i-r} < i^r \alpha^{i-r} < i \alpha^{i-1}.$$

Therefore, $i^2 \alpha^{i-1} > \sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r} > (1+\alpha)^i - \alpha^i - 1$. Hence,

$$\frac{\alpha^i}{\sum_{r=1}^{i-1} \binom{i}{r} \alpha^{i-r}} > \frac{\alpha}{i^2}.$$

Calculation 2: For $n > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^q$, $\alpha = 10^3 \ln(|\Xi|)(q+1)^2/\epsilon$, $q \geq 3$ and $i \in [q]$, for the case $i \geq 3$ we write:

$$e^{-10^{-3}\alpha^i \cdot \epsilon \cdot n^{1-(i-1)/q - i/q^2} \cdot \frac{1}{3i}} \leq e^{-\ln(|\Xi|)^3 (q+1)^6 \cdot \epsilon^{-2} \cdot n^{1-(i-1)/q - i/q^2} \cdot \frac{1}{3q}}$$

$$\leq \frac{1}{100} |\Xi|^{-q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-i/q}}$$

For $i = 1$, we use $n^{1/q - i/q^2} > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^{1-1/q} \geq 8q^2(\log(|\Xi|))^{4/3}/\epsilon^{2/3}$, and for $i = 2$ we use $n^{1/q-i/q^2} > (24q(q+1)^2(\log((q+1)|\Xi|))^2/\epsilon)^{1-2/q} \geq 2q(\log(|\Xi|))^{1/3}$. In both cases we substitute the value of α^i and write:

$$e^{-10^{-3}\alpha^i \cdot \epsilon \cdot n^{1-(i-1)/q - i/q^2} \cdot \frac{1}{3i}} = e^{-10^{-3}\alpha^i \cdot \epsilon \cdot n^{1/q - i/q^2} \cdot n^{1-i/q} \cdot \frac{1}{3i}} \leq \frac{1}{100} |\Xi|^{-q(q+1)^2 \log(|\Xi|)(\log((q+1)\log(|\Xi|)/\epsilon))^2 n^{1-i/q}}$$