Testing Union of Properties Through Universal Testing

Eldar Fischer* Oded Lachish†

November 28, 2014

Abstract

1 1-sided which doesn't work

Let $w \in \{0,1\}^n$, $i \in [n]$ and $Q \subseteq [n]$. We use w_i to denote the i'th letter of w and w_Q to be the a string $v \in \{0,1\}^{|Q|}$ such that, for every $j \in [|Q|]$, $v_j = w_{k(j)}$, where $k(j)$ is the j'th smallest member of Q.

Definition 1.1 (constraints, etc)**.** *A* q-constraint *is a pair* $C = (Q, U)$ *where* Q *is a subset of* $[n]$ *of size at most* q*, and* U *is a subset of* $\{0,1\}^{|Q|}$*. A word* $w \in \{0,1\}^n$ *is said to* violate C *if* $w_Q \in U$*.*

A q-formula F *is a set of* q*-constraints, all of whose corresponding* Q *sets are distinct. The property* P_F *is defined as the set of words that violate no member of* F*. We say that* F *is* solvable *if* $P_F \neq \emptyset$*.*

Given a property P *and a a set* $Q \subseteq [n]$*, the* natural constraint *is* $C_Q = (Q, U_Q)$*, where* $U_Q \subseteq \{0,1\}^q$ *is the set of strings* v *for which there exist no* $w \in P$ *with* $w_Q = v$*. A* witness *against a word* $w \notin P$ *is a set* Q *so that* w *does not satisfy the natural constraint* (Q, U_Q)*.*

Similarly, given a set of subsets of $[n]$*, the corresponding* natural formula *is the set of the corresponding natural constraints. It may be that such a natural formula* F *will not satisfy* $P_F = P$*, but it will always satisfy* $P_F \supseteq P$*.*

Given a word w *and a formula* F*, its* violator set $F(w)$ *is the set of the* Q*-sets of all members of* F *that are violated by* w*. In particular,* $w \in P_F$ *if and only if* $F(w) = \emptyset$*.*

Note that there can be constraints $C = (Q, U)$ where $Q = \emptyset$. In this case U is either the empty set, in which case the constraint does not restrict the input, or U contains the "null assignment" as its sole element, in which case C is violated by all words w.

Lemma 1.2 (probability quantization)**.** *Suppose that* μ *is a distribution over a set* S *of size* m*, where* $\phi \notin S$*, and let* $1 > \alpha > 0$*. There exists a distribution* μ' *over* $S' = S \cup \{\phi\}$ *satisfying the following.*

- *For every* $T \subseteq S$*,* $\mu(T) \geq \mu'(T) \geq \mu(T) - \alpha$*.*
- *The set of values* $\{\mu'(s) : s \in S\}$ *is of size at most* $2\alpha^{-1}(1 + \log m + \log \alpha) + 1$*.*

Proof. Replace all probabilities smaller than $\alpha/2m$ with 0, and all other probabilities with the lowest power of $(1 - \alpha/2)$ that is smaller than them, giving all surplus weight to the new ϕ. The second bullet uses $-\log(1 - \beta) > \beta$ for $1 > \beta > 0$ in its calculation. □

*Department of Computer Science, Technion, Haifa 32000, Israel. eldar@cs.technion.ac.il
†Birkbeck, University of London, London, UK. oded@dcs.bbk.ac.uk

Note about distributions: We usually identify a distribution μ over a formula F also with the resulting distribution over subsets of $[n]$ when taking the corresponding set of a constraint drawn by μ (recall that all constraint sets are distinct). On the direction, given a property P and a distribution μ over sets, we identify it (in the 1-sided testing context) also with the corresponding distribution over the natural formula corresponding to $\mathrm{Supp}(\mu)$.

Definition 1.3 (1-sided tests, sort of)**.** *An (ϵ, δ)-*test *for P consists is a distribution μ over subsets of $[n]$, so that for every $w \in \{0,1\}^n \setminus \mathrm{B}(P, \epsilon)$, with probability at least $1 - \delta$ a set drawn by μ would be a witness against w.*

A (q, ϵ, δ) test *is an (ϵ, δ)-test μ where $\mathrm{Supp}(\mu)$ contains only sets of size up to q.*

Theorem 1.4 (main theorem)**.** *If there exists any $(q, \epsilon/2, 1/10q)$-test for P, and n is bigger than some polynomial of $(2/\epsilon)^q$, then $\mu_{n^{-q^5}}$ is an $(\epsilon, 1/10)$-test for P.*

Definition 1.5 (subsuming set distributions)**.** *Given distributions μ and μ' over subsets of $[n]$, we say that μ* subsumes *μ' if there exists a distribution $\tilde{\mu}$ over pairs of subsets of $[n]$ satisfying the following.*

- *For every $(A, B) \in \mathrm{Supp}(\tilde{\mu})$ we have $A \supseteq B$.*
- *The projection on the first coordinate of $\tilde{\mu}$ is μ.*
- *The projection on the second coordinate of $\tilde{\mu}$ is μ'.*

*Given μ and μ', we say that μ η-*subsumes *μ' if there exists a distribution μ'' that is η-close to μ' and is subsumed by μ.*

Lemma 1.6. *The subsuming relation is transitive. Moreover, if μ α-subsumes μ' and μ' β-subsumes μ'' then μ $(\alpha + \beta)$-subsumes μ''.*

Proof. For the first part, define $\hat{\mu}(A, B, C) = \tilde{\mu}(A, B) \cdot \tilde{\mu}'(B, C)$, and then project it on the first and third coordinate. The second part is similar with a bit more massaging. □

Observation 1.7. *If μ' subsumes μ, then for any $A \in \mathrm{Supp}(\mu')$, the probability that A is contained in a set drawn according to μ' is at least $\mu(A)$.*

Definition 1.8 (sunflower)**.** *A t-*sunflower with core *A is a family of subsets $B_1, \ldots, B_t \subseteq \{1, \ldots, n\}$ so that every B_i contains A, and $B_1, \ldots, B_t$ are disjoint outside of A (a completely disjoint family is a sunflower with core $A = \emptyset$).*

Lemma 1.9 (sunflower theorem, Erdös and Rado [1])**.** *Any family of at least $s = q! t^{q+1}$ sets whose sizes are at most q contains a sub-family of size t which is a sunflower.*

Lemma 1.10. *If F is a solvable formula, μ a distribution over F that is (ϵ, δ)-averse to to a word w, and $F' \subseteq F$ is a sub-formula whose constraints cover less than ϵn many indexes, then $\mu(F') < \delta$.*

Lemma 1.11 (main mechanism)**.** *If F is a solvable q-formula, μ is a distribution over F that is $(\epsilon, 1/10q)$-averse to a word w, $\eta < \epsilon$ and n is bigger than some global polynomial of $(2/\eta)^q$, then there exists a $(q-1)$-formula F' (not necessarily solvable) and a distribution μ' over F' satisfying the following.*

- *μ' is $(\epsilon - \eta, 1/10(q-1))$-averse to w.*

- *With probability at least* $1 - 1/10q^2$*, a set* U *drawn according to* $\mu_{n^{-q^4}}$ *is such that any word* w' *that satisfies* P_F *and agrees with* w *over* U *has to satisfy* $P_{F'}$*. In particular, if* $P_{F'}$ *is not solvable then this is the probability that* U *is a witness against* P_F*.*

Before we prove the above...

Definition 1.12 (cleaned distribution)**.** *Given a distribution over members of a solvable* q*-formula* F*, and a word* w*, we perform the following in order (please add parameters...).*

1. *Move to a quantized distribution* μ'*.*
2. *Greedily take equal-probability sunflowers until none are left.*
3. *Mark leftover sets for ignoring.*
4. *In every sunflower, consider only possible the assignments to the core to which at least* $1/10q2^q$ *of the petals "raise an objection" through* w*. Mark all petals not raising objections against any such assignment for ignoring.*

The cleaned distribution $\tilde{\mu}$ *is* μ' *when conditioned on the event of not choosing any of the sets marked for ignoring.*

Lemma 1.13. *Moving from* $(\epsilon, 1/10q)$*-averse distributions to* $(\epsilon - \eta, 1/(10q-5))$*-averse ones.*

Proof. Please put one. First step has Lemma 1.2. Third step because of Lemma 1.10 following Lemma 1.9. The ignored petals in the last step take a small probability because they are a small part of every uniform-distribution sunflower (they could cover lots of indexes, who cares). □

Proof of Lemma 1.11. First move to a cleaned distribution $\tilde{\mu}$. Then define F' according to the cores of the sunflowers in the cleaned distribution. For every core of a sunflower, forbid all assignments that are forbidden by (non-ignored) petals. If a core appears in more than one sunflower (could be for different set-probabilities ones), "unionize" the constraints. The distribution μ' chooses a core with probability the sum of probabilities of petals of the corresponding sunflower (or sunflowers).

Now prove that with probability at least $1-1/10q^2$ it happens that a set U chosen according to $\mu_{n^{-q^4}}$ contains a full "witness petal" for every forbidden assignment of every constraint in F'. Given this event, the second item of the lemma clearly holds.

Now prove that the union of the set Q of a constraint C drawn according to μ', with a set drawn according to $\mu_{n^{-q^4}}$, $1/10q^2$-subsumes $\tilde{\mu}$ when considered as a distribution over the sets of F. This is since when the above event happens, you can first choose a core according to μ' and then uniformly choose among the petals contained in U.

It is still a problem to prove the first part of the lemma... □

2 2-sided which is not yet known not to work

Definition 2.1 (probabilistic constraints and formulas)**.** *A* probabilistic q-constraint *is a pair* $C = (Q, S)$ *where* $Q \subseteq [n]$ *is a* constraint set *of size at most* q *and* S *is a* satisfaction function *from* $2^{|Q|}$ *to the real interval* $[0, 1]$*.*

A probabilistic q-formula $P = (F, \mu)$ *is a set* P *of* q*-constraints, all with distinct constraint sets, along with probability distribution* μ *over* F*.*

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

Given a word w and a probabilistic formula P, the satisfaction *of P by w is the average of the random variable of picking a constraint $(Q, S) \in F$ according to P and obtaining the value $S(w_Q)$. P is said to be* α-sure for w *if its satisfaction is outside the range $(\alpha, 1-\alpha)$.*

The satisfiability *of P is the maximum satisfaction of P among all possible words w.*

Definition 2.2 (2-sided non-adaptive test)**.** *A* 2-sided (ϵ, δ, q)-test *for a property L is a probabilistic q-formula so that its satisfaction is at least $1-\delta$ by any word in L, while its satisfaction is at most δ by any word ϵ-far from L.*

We henceforth drop the word "probabilistic" when talking about constraints and formulas. Note that it is possible for a constraint $C = (Q, S)$ to have $Q = \emptyset$, in which case S is of the type $\{\emptyset \longmapsto \alpha\}$, that is, the constraint has a satisfaction value that is not dependent on w.

Lemma 2.3. *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 $P = (F, \mu)$, then we define 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. □

Observation 2.4. *If $P = (F, \mu)$ is $\frac{1}{r}$-sure for w where r is some integer, and $F' \subseteq F$ satisfied $\mu(F \setminus F') \leq s/r$, then the F'-conditioned formula P' is $\frac{1}{r-s}$-sure for w.*

Whenever F is clear from context (and satisfies the distinct constraint set requirement), we freely identify the constraints with their corresponding sets. In particular distributions over P are identified with distributions over the corresponding subsets of $[n]$, and sunflowers of members of F are sunflowers of the corresponding sets with the additional requirement that μ is constant over the participating members of P.

Definition 2.5 (compound constraints and formulas)**.** *If $C_1, \ldots, C_t$ are constraints forming a sunflower with core A, and given a word w, the* compound constraint *C has the constraint set A, and the satisfaction function defined by $S(v) = \frac{1}{t} \sum_{i=1}^{t} S_i(v \sqcup w_{Q_i \setminus A})$.*

Given a formula P where the constraint set F is divided to sunflowers $F_1, \ldots, F_r$ with cores $A_1, \ldots, A_r$ respectively, the compound formula *$P' = (F', \mu')$ is defined by F' consisting of the compound constraints corresponding to every sunflower, and the distribution μ' defined so that for a compound constraint C, its probability is the sum of probabilities given by μ to the members of the original sunflower (if some flowers have the same center, we use Lemma 2.3).*

Given P as above and a constraint subset $H \subseteq F$, the H-approximated compound formula *is defined by F' consisting of the compound constraints corresponding to the intersections $F_1 \cap H, \ldots, F_r \cap H$, and μ' being the same as that of the compound formula for P (if $F_j \cap H = \emptyset$ for some j then we arbitrarily set the corresponding satisfaction function to be a constant $\frac{1}{2}$).*

It is important to note that a compound constraint is calculated independently of the value of the word w inside the core A itself.

Definition 2.6 (formula representation)**.** *Given two formulas F and F', we say that F'* α-represents *F around a word w, if for every word w' that differs from w in at most s places (for any t), the satisfaction of F and F' by w' differ by at most αs.*

We say that F' (α, β)-represents *F around w, if for every w' in the above setting the satisfaction of F and F' by w' differ by at most $\alpha s + \beta$.*

Observation 2.7. *If F' (α, β)-represents F around w and F'' (α', β')-represents F' around w, then F'' $(\alpha + \alpha', \beta + \beta')$-represents F around w.*

Lemma 2.8. *If P' is the compound formula of P according to a word w, then the satisfaction of P and P' by w is the same. Moreover, if the sunflowers $F_1, \ldots, F_r$ used for the compound formula P' are all of size at least t, and w' is obtained from w by changing at most s letters, then the satisfaction of P and the satisfaction of P' by w' differ by no more than s/t.*

Proof. The first statement is chain rule etc. For the second statement construct the compound formula $\tilde{P}$ of P according to w', and then note that each of its (compound) constraints has a satisfaction function differing by not more than s/t (in l_∞ norm) from that of the corresponding compound constraint of P'. □

Lemma 2.9 (large deviation bound)**.** *Suppose that $\alpha_1, \ldots, \alpha_m$ are all values in $[0,1]$, and let $U \subseteq [m]$ be chosen according to μ_p for $p \geq 10c/\eta^2 m$ where $c > 1$. Then with probability at least $1 - 2^{-c}$ the value $(\sum_{i \in U} \alpha_i)/|U|$ (where we arbitrarily set it to $\frac{1}{2}$ if $U = \emptyset$) is in the range $(\sum_{i=1}^m \alpha_i)/m \pm \eta$.*

Proof. Surely there must be a nice proof and with the correct coefficients somewhere... I can think of a kludgy one that starts with bounding the size of U and then for each fix size use that selecting without repetitions is no less concentrated than with repetitions, and then a standard deviation inequality. □

Lemma 2.10. *Suppose that a q-formula F is partitioned to sunflowers of size at least t. Let $U \subseteq [n]$ be chosen according to μ_p where $p = (2000q^5 4^q \log n/t)^{1/q}$, and let H be the set of constraints whose "petal part" is contained in U. That is, denoting the sunflowers by $F_1, \ldots, F_r$ and their cores by $A_1, \ldots, A_r$, we set $H = \bigcup_{j=1}^r \{(Q,S) : (Q,S) \in S_j \wedge Q \setminus A_j \subseteq U\}$. With probability at least $1/3q$, the H-approximated compound formula of F (with respect to a given word w) is a $(1/t, 1/10q^2)$-representation of F around w.*

Proof. We compare the resulting H-approximated compound formula with the corresponding (non-approximated) compound formula. For every sunflower F_j with center A_j and every possible assignment $v \in \{0,1\}^{|A_j|}$ we compare the corresponding value of the satisfaction function of the compound constraint $C_j = (A_j, S_j)$ and the approximated compound constraint $C'_j = (A_j, S'_j)$, noting that inside the sunflower the the distribution of $H \cap F_j$ is just like invoking $\mu_{p^{q-|A_j|}}$ over F_j. We use Lemma 2.9 with $\eta = 1/10q^2 2^q$ and $c = \log(3qn^q 2^q)$. A union bound over every value of every satisfaction function of every constraint of every sunflower brings us to a situation where the approximated compound formula $(0, 1/10q^2)$-represents the compound formula, which in turn by Lemma 2.8 $(1/t, 0)$-represents F around w, allowing us to conclude by Observation 2.7. □

Observation 2.11 (multi-round sampling)**.** *An algorithm making k rounds of sampling using $\mu_{p_1}, \ldots, \mu_{p_k}$ respectively can be converted to an algorithm making one round of sampling μ_p for $p = 1 - \prod_{i=1}^k (1 - p_i)$*

Proof. First note that p is the probability for any given index to be picked in at least one of the rounds. To convert the algorithm, first choose U according to μ_p, and then for every index $j \in U$ say that it appears in exactly the rounds I_j (where $\emptyset \neq I \subseteq [k]$) with probability $(\prod_{i \in I} p_i)(\prod_{i \notin I}(1 - p_i))/p$. Now run the original algorithm setting $U_i = \{j : i \in I_j\}$ as the set chosen in round i for every i, noting that all probabilities stay the same. □

Theorem 2.12. *A 2-sided $(\epsilon/2, 1/10q, q)$-test can be converted to an ϵ-test whose querying method consists of just sampling according to μ_p for $p = n^{1-\alpha(q)}$, where $\alpha(q)$ is some global positive function, as long as n is larger than some global polynomial of $1/\epsilon$ and $\alpha(q)$.*

Proof. We will do the sampling in q rounds, considering Observation 2.11. The probabilities $p_1, \ldots, p_q$ will also have coefficients depending on q and $\log n$, but such coefficients are subsumed (decreasing somewhat the power of n) by the assumption that n is large enough. We Define $t_1, \ldots, t_q$ along the way, and our sampling probabilities will be those related to Lemma 2.10 (noting that $(q-i)$-formulas are also q-formulas, so the lemma will work in all rounds). $p_1, \ldots, p_q$ will be the corresponding probabilities provided by the lemma, while $t_1, \ldots, t_q$ will be powers of n with some coefficients polynomial in q and $\log n$.

The analysis is such that the only possibility for an error is when the small probability bad event of Lemma 2.10 happens, and for q rounds the probability of it happening throughout the algorithm is bounded by $\frac{1}{3}$. The rest of the analysis is assuming that it didn't happen.

We assume that w is either a word satisfying the property or a word $\epsilon/2$-far from satisfying it, so in particular P is originally $1/10q$-sure for w. At every round of the algorithm, we either calculate a scheme for moving to a compound formula or gain sufficient knowledge on the satisfaction of the current formula to stop and provide an answer. In the first case we then use sampling and calculate an approximated compound formula, in particular a $(q-i)$-formula, for the next round. If we didn't stop earlier, after q rounds we are left with a 0-formula, which is basically just one satisfaction value, so we can surely stop then.

Formally we start the i round with a $(q-i+1)$-formula the is $1/10(q-i+1)$-sure for w, and either output an answer or end the round with a $(q-i)$-formula that is $1/10(q-i)$-sure for w. In the case of w being ϵ-far from the property there will be a stronger invariant in that the resulting formula will be $1/10(q-i)$-sure in rejecting any word that differs from w in less than $\min\{\epsilon n/2, t_i/10q^2\}$ places. The first value is subsumed in the second for any n large enough, so we will henceforth ignore it.

Denoting the formula from the $i-1$ round by P_{i-1}, to construct P_i we start by quantizing P_{i-1} by Lemma 1.2 to get a $1/(10(q-i)+9)$-sure formula P'_{i-1} that has at most $500q^2\log(qn)$ probability values. Now we define t_i as the largest value so that $q!(t_i)^{q+1} \cdot 500q^2\log(qn) \cdot q < t_{i-1}/10q^2$. We then greedily take sunflowers (every sunflower of equal probability constraints) until we are left with at most $q!(t_i)^{q+1} \cdot 500q^2$ constraints of each of the possible $500q^2\log(qn)$ probability values, in particular covering a total of less than $t_{i-1}/10q^2$ indexes. P''_{i-1} be the conditioning of P'_{i-1} on not picking one of the remaining constraints outside the sunflowers, while Q_{i-1} be the conditioning of P'_{i-1} on only picking constraints outside the sunflowers.

Now there are two cases. The first case is where the total weight of constraints outside the sunflowers is at most $3/(10(q-i)+9)$. In this case P''_{i-1} is $1/(10(q-i)+6)$-sure for w, and its corresponding compound formula $\hat{P}_i$ is also $1/(10(q-i)+6)$-sure for w. Additionally, if w is ϵ-far from the property and w' differs from w in less than $t_i/10q^2$ places then $\hat{P}_i$ is still $1/(10(q-i)+5)$-sure about rejecting w' by Lemma 2.8. To construct P_i as required we do a sampling with the p_i as per Lemma 2.10 and construct the corresponding approximate compound formula.

The second case is when the weight of the constraints outside the sunflowers of P'_{i-1} is more than $3/(10(q-i)+9)$. In this case we calculate the maximum satisfiability of Q_i (without any additional sampling), and act accordingly. If it is at least $\frac{1}{2}$ then we accept the input w, and otherwise we reject it. We claim that this decision is not wrong. If we accepted the input, then it means that we can change w in less than $t_{i-1}/10q^2$ places and reach a word w' for which P'_{i-1} cannot be $1/(10(q-i)+9)$-sure about rejecting, and so it cannot be the situation that w was ϵ-far from the property. If we rejected the input, then no word at all (including w itself) can cause P'_{i-1} to be $1/(10(q-i)+9)$-sure about accepting, and so w cannot be a word in the property. □

References

[1] Paul Erdős and Richard Rado. Intersection theorems for systems of sets. *J. London Math. Soc.*, 35:85–90, 1960.