On properties that are non-trivial to test

Nader H. Bshouty* Oded Goldreich†

December 19, 2024

Abstract

In this note we show that all sets that are neither finite nor too dense are non-trivial to test in the sense that, for every $\epsilon > 0$, distinguishing between strings in the set and strings that are ϵ-far from the set requires $\Omega(1/\epsilon)$ queries. Specifically, we show that if, for infinitely many n's, the set contains at least one n-bit long string and at most $2^{n-\Omega(n)}$ many n-bit strings, then it is non-trivial to test.

A preliminary version of this work was posted as TR22-013 of *ECCC*.

The main result. This note refers to the query complexity of property testing (see the textbook [2]). Specifically, a tester for a set of strings S is explicitly given two parameters, a length parameter $n \in \mathbb{N}$ and a proximity parameter $\epsilon > 0$, as well as query access to an n-bit string x. The tester is required to distinguish the case that x is in S from the case that x is ϵ-far from S, where x is ϵ-far from S if its Hamming distance from each $|x|$-bit long string in S is greater than $\epsilon \cdot |x|$. (By distinguishing between strings in A and strings in B we mean accepting each string in A with probability at least 2/3 and rejecting each string in B with probability at least 2/3.)

Definition 1 (non-trivial to test): *A set of strings S is* non-trivial to test *if, for every $\epsilon > 0$ and infinitely many $n \in \mathbb{N}$, the query complexity of testing S, with parameters n and ϵ, is $\Omega(1/\epsilon)$.*

Theorem 2 (sufficient condition for non-triviliaty): *Suppose that, for infinitely many n's, the set S contains at least one n-bit long string and at most $2^{n-\Omega(n)}$ many n-bit strings. Then, S is non-trivial to test.*

Note that the sufficient condition is necessary in general. In particular, a set S that, for every n, contains $2^{n-o(n)}$ many n-bit long strings *may* be trivial to test in the sense that, for every $\epsilon > 0$ and all sufficiently large n, every n-bit long string is ϵ-close to S.

Proof: We use a reduction from the special case in which every n-bit long string in S has Hamming weight at most $n - \Omega(n)$. Letting w be an n-bit long string of maximum Hamming weight, we consider a random variable X obtained from w by flipping each 0-entry in w to 1 with probability

*Department of Computer Science, Technion, Haifa, Israel. Email: `bshouty@cs.technion.ac.il`.

†Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. Email: `oded.goldreich@weizmann.ac.il`. Partially supported by the Israel Science Foundation (grant No. 1146/18). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702).

$O(\epsilon)$. We observe that X is ϵ-far from S and that distinguishing w from X requires $\Omega(1/\epsilon)$ queries. Transforming each instance of the general case to an instance of the special case (by XORing with a random string) we establish the theorem. Details follow.

Let $c<1$ be a constant such that for infinitely many n's the set $S^{(n)}=S\cap\{0,1\}^n$ is non-empty and contains at most 2^{cn} strings. For a sufficiently small $\eta=\eta(c)>0$, we shall first show that for such n's there exists $r\in S^{(n)}$ such that the relative Hamming weight of each string in $r\oplus S^{(n)}=\{r\oplus s: s\in S^{(n)}\}$ is at most $1-\eta$.

The foregoing claim is proved by the probabilistic method. Letting $\mathtt{wt}(x)=|\{i\in[|x|]: x_i=1\}|/|x|$ denote the relative Hamming weight of x, we have

$$\begin{aligned}\Pr_{r\in\{0,1\}^n}\left[\exists s\in S^{(n)}\ \ \mathtt{wt}(r\oplus s)>1-\eta\right] &\le |S^{(n)}|\cdot\Pr_{r\in\{0,1\}^n}[\mathtt{wt}(r)>1-\eta]\\ &\le 2^{c\cdot n}\cdot\sum_{i<\eta n}\binom{n}{i}\cdot 2^{-n}\\ &= 2^{(c+H_2(\eta)-1)\cdot n} < 1,\end{aligned}$$

where H_2 denotes the binary entropy function. Hence, there exists an n-bit string r such that $\tau\stackrel{\rm def}{=}\max_{s\in S^{(n)}}\{\mathtt{wt}(r\oplus s)\}\le 1-\eta$, and let $w\in r\oplus S^{(n)}$ be such that $\mathtt{wt}(w)=\tau$.

For every $\epsilon\in(0,\eta/2)$, let X be a random variable, distributed over n-bit strings, such that if $w_i=1$ then $X_i=1$ and otherwise $\Pr[X_i=1]=2\epsilon/\eta$ independently of all other X_j's. Note that $\mathrm{E}[\mathtt{wt}(X)]=\mathtt{wt}(w)+\frac{2\epsilon}{\eta}\cdot(1-\mathtt{wt}(w))\ge\mathtt{wt}(w)+2\epsilon$ (equiv., $\mathrm{E}[\sum_{i:w_i=0}X_i]=2\epsilon\cdot n$). Hence, assuming $n=\omega(1/\epsilon)$, with high probability, X is ϵ-far from $r\oplus S^{(n)}$, since $\Pr[\mathtt{wt}(X)>\mathtt{wt}(w)+\epsilon]=1-o(1)$ (equiv., $\Pr[\sum_{i:w_i=0}X_i>\epsilon n]=1-o(1)$), whereas $\max_{s\in S^{(n)}}\{\mathtt{wt}(r\oplus s)\}=\mathtt{wt}(w)$. On the other hand, distinguishing $w\in r\oplus S^{(n)}$ from X requires $\Omega(\eta/\epsilon)=\Omega(1/\epsilon)$ queries, since $\Pr[X_i\neq w_i]\le 2\epsilon/\eta$ for every $i\in[n]$.

It follows that ϵ-testing $r\oplus S^{(n)}$ (i.e., distinguishing strings in $r\oplus S^{(n)}$ from strings that are ϵ-far from $r\oplus S^{(n)}$) requires $\Omega(1/\epsilon)$ queries. The theorem follows, since ϵ-testing $r\oplus S^{(n)}$ reduces to ϵ-testing $S^{(n)}$ (i.e., given an ϵ-tester for $S^{(n)}$, we obtain an ϵ-tester for $r\oplus S^{(n)}$ by XORing the input string with r (and observing that the distance of x from $r\oplus S^{(n)}$ equals the distance of $x\oplus r$ from $S^{(n)}$)). ■

Digest. A key observation used in the proof is that shifting a (not too dense) set by XORing its elements with a random string yields a set of strings such that each string has relative Hamming weight that is closed to 0.5. Observing that the pairwise distances between strings is preserved and replacing η by $0.5-\epsilon$, we obtain the following result (where n and $k=k(n)$ are viewed as varying).

Proposition 3 (obtaining almost balanced error correcting codes): *Let $C:\{0,1\}^k\to\{0,1\}^n$ be an error correcting code of relative distance δ, and ϵ be such that $\frac{k}{n}+H_2(0.5-\epsilon)$ is upper-bounded by a constant that is smaller than 1. Then, with very high probability over the choice of $r\in\{0,1\}^n$, it holds that $C_r:\{0,1\}^k\to\{0,1\}^n$ such that $C_r(x)=C(x)\oplus r$ is an error correcting code of relative distance δ in which all codewords have relative Hamming weight $0.5\pm\epsilon$.*

Proof: Analogously to the proof of Theorem 2, we have

$$\begin{aligned}\Pr_{r\in\{0,1\}^n}\left[\exists x\in\{0,1\}^k\ \ \mathtt{wt}(r\oplus C(x))\notin[0.5\pm\epsilon]\right] &\le 2\cdot 2^k\cdot\sum_{i<(0.5-\epsilon)\cdot n}\binom{n}{i}\cdot 2^{-n}\\ &= 2^{1+\left(\frac{k}{n}+H_2(0.5-\epsilon)-1\right)\cdot n}\end{aligned}$$

and the claim follows by the hypothesis that $\frac{k}{n} + H_2(0.5 - \epsilon)$ is upper-bounded by a constant that is smaller than 1. ■

Postscript. Subsequent to our work, Fischer [1] proved a stronger result of a similar nature. Specifically, both papers yield an $\Omega(1/\epsilon)$ query lower and both assume that the property of n-bit strings is non-empty, but we assume that the property has at most $2^{n-\Omega(n)}$ strings, whereas Fischer assumes the existence of a string that is $\Omega(1)$-far from the property. Note that our hypothesis imply Fischer's (i.e., if each n-bit string is α-close to $S^{(n)}$, then $\frac{|S^{(n)}|}{2^n} \geq 2^{-H_2(\alpha)\cdot n}$). On the other hand, Fischer's result can be proved by following our proof strategy.[1]

Acknowledgements. We thank Rocco Servedio for a useful discussion. We are also grateful to Augusto Modanese for pointing out an unclear argument in a previous verion.

References

[1] E. Fischer. A Basic Lower Bound for Property Testing. `arXiv:2403.04999 [cs.DS]`.

[2] O. Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[1] Specifically, if $s \in \{0,1\}^n$ is α-far from $S^{(n)}$, then 0^n is α-far from $S^{(0)} \oplus s$. Switching between 0s and 1s, we obtain a non-empty set such that all strings in it have maximal weight $1 - \alpha$, and proceed as in the last part of the proof of Theorem 2.