WoLA'19: Open Problems

July, 2019

Abstract

Some questions suggested during the Open Problems session of the 3rd Workshop on Local Algorithms (WoLA), held in July 2019 at ETH, Zurich.

Non-Adaptive Group Testing Suggested by Oliver Gebhard.

In (non-adaptive) quantitative group testing, one has a population of n individuals, among which $k = n^c$ (for some constant $c \in (0,1)$) are sick. The goal is, by performing m non-adaptive tests, to identity the k sick individuals (where a test is a subset $S \subseteq [n]$, whose output is 1 if S contains at least one sick individual).

By a counting argument, one gets a lower bound of $m = \Omega\Big(\frac{k}{\log k} \log \frac{n}{k}\Big)$ tests; however, the best known upper bound is $m = O(k \log \frac{n}{k})$.

Question 1. Can one get rid of the $\log k$ factor in the lower bound; or, conversely, improve the upper bound to match it?

Distribution Testing: identity testing up to coarsenings Suggested by Clément Canonne.

Given a distance parameter $\varepsilon \in (0,1]$, i.i.d. samples from an unknown distribution p and a (known) reference distribution q, both over $[n] = \{1, \dots, n\}$, the identity testing question asks for the minimum number of samples sufficient to distinguish, with probability at least $2/3$, between (i) $p = q$ and (ii) $\mathrm{d_{TV}}(p,q) > \varepsilon$. ($\mathrm{d_{TV}}$ here denotes the total variation distance.) This question is by now fully resolved, with $\Theta(\sqrt{n}/\varepsilon^2)$ samples being necessary and sufficient [1, 2].

However, consider the following variant: given a (fixed) family $\mathcal{F}$ of functions from $[n]$ to $[m]$, and a reference distribution q over $[m]$, distinguish between (i) there exists $f \in \mathcal{F}$, $p = q \circ f$, and (ii) $\min_{f \in \mathcal{F}} \mathrm{d_{TV}}(p, q \circ f) > \varepsilon$.

This $\mathcal{F}$*-identity testing* question includes the identity testing one as special case by setting $m = n$ and $\mathcal{F}$ to be the singleton containing the identity function. One can also take $m = n$ and $\mathcal{F}$ to be the class of all permutations, to test "identity up to relabeling" (a problem whose sample complexity is, from previous work of Valiant and Valiant, known to be $\Theta(n/(\varepsilon^2 \log n))$ (see [3, Corollary 11.30]).

Question 2. For a fixed m, and $\mathcal{F}$ the family of all partitions of $[n]$ into m consecutive intervals, what is the sample complexity of $\mathcal{F}$-identity testing, as a function of n, ε, m?

Note: this corresponds to testing whether p is a "refinement" of the coarse distribution q; or, equivalently, if p ad q are the same, up to the precision of the measurements.

LCA for MIS Suggested by Mohsen Gaffhari.

In the model of Local Computation Algorithms (LCA), given an input graph $G = (V, E)$, an algorithm gets, upon query any vertex v of its choosing, the list of neighbors of v. In this model, the current state-of-the-art for the query complexity of computing a Maximal Independent Set (MIS) for graph G of maximum degree at most Δ is an upper bound of $\Delta^{O(\log\log\Delta)} \operatorname{polylog} n$ queries.

Question 3. Does there exist a $\operatorname{poly}(\log n, \Delta)$-query LCA for MIS?

Estimating a graph's degree distribution Suggested by C. Seshadhri.

The *degree distribution* of a graph $G = (V, E)$ is the histogram of the degree frequencies: i.e., letting $n(d)$ denote the number of degree-d vertices, the histogram $(n(d))_{d\geq 0}$. Define the (complementary) cumulative distribution function as

$$N(d) \stackrel{\text{def}}{=} \sum_{d'\geq d} n(d'), \qquad d \geq 0\,.$$

Assume one has access to the graph G via the following three types of queries:

1. sampling a u.a.r. vertex
2. querying the degree of a given vertex
3. sample a u.a.r. neighbor of a given vertex

and the goal is to obtain the following $(1 \pm \varepsilon)$-"bicriteria" approximation $\hat{N}$ of the degree distribution: for all d,

$$(1-\varepsilon)N((1-\varepsilon)d) \leq \hat{N}(d) \leq (1+\varepsilon)N((1+\varepsilon)d)\,.$$

Previous work of Eden, Jain, Pinar, Ron, and Seshadhri [4] shows an upper bound of

$$\frac{n}{h} + \frac{m}{\min_d d \cdot N(d)}$$

queries, where h is the value s.t. $N(h) = h$ (where the complementary cdf intersects the diagonal).

Question 4. Can this upper bound be improved? Can one establish matching lower bounds?

And also, slightly less well-defined:

Question 5. Can one obtain better upper bounds when relaxing the goal to only learn the *high-degree* (tail) part of the distribution? What about testing properties of the degree distribution (e.g., "power-law-ness") in this setting? And what about the first type of queries – can one relax it, or work with a different type of sampling than uniform (for instance, via random walks)?

About the uniform vertex sampling Suggested by Oded Goldreich.

The graph query model where one gets to query vertices uniformly at random, as mentioned in the previous open problem, may seem unrealistic in some cases. Thus, one may advocate alternative models, especially in the context of graph property testing, akin to the "distribution-free" model of property testing (for functions) and the PAC model (for learning). In this *Vertex-Distribution-Free* (VDF) model of testing suggested in a recent paper [5],[1] one gets i.i.d. vertices sampled from an

[1] This model was briefly discussed in [6, Section 10.1].

arbitrary distribution $\mathcal{D}$ over the vertex set, and the goal is to test w.r.t. to the (pseudo) distance induced by $\mathcal{D}$.

Question 6. Perform a systematic study of property testing, both in the bounded-degree and dense graph models, in this VDF setting.

Question 7 (Suggested by C. Seshadhri). Can one define, motivate, and prove non-trivial results in an *Edge*-Distribution-Free model, analogous to the VDF one but with regard to sampling random edges?[2]

Effective support size estimation in the dual model Suggested by Oded Goldreich.
For a probability distribution p over a discrete domain Ω, and a parameter $\varepsilon \in [0,1]$, denote by

$$\mathrm{ess}_\varepsilon(p) \stackrel{\text{def}}{=} \min\{\mathrm{supp}(q) : \mathrm{d_{TV}}(p,q) \le \varepsilon\}$$

the ε-effective suport size of p, i.e., the smallest possible support size of any distribution ε-close to p. This turns out to be a more robust and interesting measure in general than the support of p, which is $\mathrm{ess}_0(p) = \mathrm{supp}(p)$. In recent work, Goldreich [7] focused on the query complexity of approximating the effective support size of a discrete distribution provided via two oracles: sampling (samp_p), and query access (to the probability mass function), eval_p. In particular, the goal is, given parameters ε and $\beta > 1$, to output an $f(\varepsilon, \beta, n)$-factor approximation of $\mathrm{ess}_{\varepsilon'}(p)$, for some $\varepsilon' \in [\varepsilon, \beta\varepsilon]$.
In the aforementioned work, algorithms are obtained achieving (for constant $\beta > 1$)

- query complexity $\mathrm{poly}(1/\varepsilon)$ and approximation factor $f = O(\log\log\log\log(n/\varepsilon))$, that is, any constant number of iterated logarithms;
- query complexity $\mathrm{poly}(\log^* n, 1/\varepsilon)$ even for approximation factor $f = O(1)$;

where $n \stackrel{\text{def}}{=} \mathrm{ess}_\varepsilon(p)$. (As well as several other results interpolating between the two extremes.)

Question 8. Can one get the best of both worlds, and get rid of the $\log^* n$ to obtain query complexity $\mathrm{poly}(1/\varepsilon)$ *and* constant approximation factor?

Vertex connectivity in the LOCAL model Suggested by Sorrachai Yingchareonthawornchai.
In this question, the input is the underlying graph $G = (V, E)$, as well as parameters ν, k and vertex $v \in V$. The goal is to output either $\perp$ or a subset $S \subseteq V$, such that

- if $\perp$ is the output, there is no S such that $v \in S$ with $|S| \le \nu$ and $|N(S)| < k$;
- if the output is a set S, then $|N(S)| < k$.

It is known that this problem can be solved with $O(\nu k)$ queries, and either time $O(\nu^{3/2}k)$ (deterministic) or $O(\nu k^2)$ (randomized) [8, 9, 10].

Question 9. Can one achieve time $O(\nu k)$?

Making edges happy in the LOCAL model Suggested by Jukka Suomela.
In this question, the input is the underlying graph $G = (V, E)$, promised to have maximum degree at most Δ, and the goal is to compute an orientation of the edges of E which makes all edges

[2] This type of variant was also briefly evoked in [6, Section 10.1.4], where it was shown that Bipartiteness is not testable in such an EDF model.

"happy." Specificaly, for any given orientation of the edges, the *load* of a node $v \in V$ is its number of incoming edges. An edge e is then said to be *happy* if switching its orientation does not make it point to a smaller-node load.

One can show by a greedy argument that there always exists an orientation making all edges happy. Moreover, a surprising result established that, in the LOCAL model, such a configuration could be found in $\mathrm{poly}(\Delta)$ rounds, *independent* of the number of nodes n. However, the question of the dependence on Δ remains wide open, as even a $\mathrm{polylog}(\Delta)$ upper bound is not ruled out.

Question 10. What is the right dependence on Δ? Can one show *any* lower polynomial lower bound, e.g., $\Delta^{0.1}$, $\sqrt{\Delta}$, or Δ?

References

[1] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Information Theory*, 54(10):4750–4755, 2008. (document)

[2] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. (document)

[3] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. (document)

[4] Talya Eden, Shweta Jain, Ali Pinar, Dana Ron, and C. Seshadhri. Provable and practical approximations for the degree distribution using sublinear graph samples. In *WWW*, pages 449–458. ACM, 2018. (document)

[5] Oded Goldreich. Testing graphs in vertex-distribution-free models. In *STOC*, pages 527–534. ACM, 2019. (document)

[6] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998. 1, 2

[7] Oded Goldreich. On the complexity of estimating the effective support size. *Electronic Colloquium on Computational Complexity (ECCC)*, 26:88, 2019. (document)

[8] Danupon Nanongkai, Thatchaphol Saranurak, and Sorrachai Yingchareonthawornchai. Breaking quadratic time for small vertex connectivity and an approximation scheme. In *STOC*, pages 241–252. ACM, 2019. (document)

[9] Danupon Nanongkai, Thatchaphol Saranurak, and Sorrachai Yingchareonthawornchai. Computing and testing small vertex connectivity in near-linear time and queries. *CoRR*, abs/1905.05329, 2019. (document)

[10] Sebastian Forster and Liu Yang. A faster local algorithm for detecting bounded-size cuts with applications to higher-connectivity problems. *CoRR*, abs/1904.08382, 2019. (document)