Generalized Uniformity Testing

Tuğkan Batu and Clément Canonne
Broader Picture: Inferring from Data

Big datasets and/or continuous stream of data: need to check quickly if some property of interest holds.

- See the dataset as a probability distribution
- Connection to hypothesis testing, model selection
- Formalism: property testing of distributions [GR00,BFR+00]

Narrower Picture: Distribution Testing

- **Discrete** domain $\Omega$
- **Fixed property** of distributions $\mathcal{C} \subseteq \Delta(\Omega)$
- Access to i.i.d. samples from **unknown, arbitrary** distribution $p$
- Distance parameter $\varepsilon \in (0,1)$

**Must decide**

$$p \in \mathcal{C} \text{ vs. } \text{TV}(p, \mathcal{C}) > \varepsilon$$

(with probability $\frac{2}{3}$)
Distribution Testing
A lot of (tight) results results in testing of discrete distributions over known domain $\Omega=\{1,...,n\}$: uniformity, identity, closeness, independence, monotonicity, log-concavity, juntas, MHR, PBD, SIIRV, histograms,... [BFF+01, BKR04, Pan08, LRR11, VV14, ADK15, DKN15, BFR+10, CDVV14, Can16, DK16, DKS17,...]

Let’s focus on uniformity.
Uniformity Testing

Given samples from an arbitrary $p \in \Delta(\Omega)$, distinguish $p = u_\Omega$ from $\text{TV}(p, u_\Omega) > \varepsilon$.

**First, fundamental testing question.**

[GR00], [BFR+00], [Pan08], [DKN15], [DGPP16]

$\Theta(\sqrt{|\Omega|}/\varepsilon^2)$

**Catch:** For known domain $\Omega$. 
Generalized Uniformity Testing

Given samples from an arbitrary $p \in \Delta(\Omega)$, distinguish $p = u_\Omega$ from $\text{TV}(p, u_\Omega) > \varepsilon$.

But we do not know $\Omega$. (Why would we?)

So... still $\Theta(\sqrt{|\Omega|/\varepsilon^2})$?

Answer: “No.”
Generalized Uniformity Testing

“You get samples from a discrete unknown set. Is the underlying distribution uniform?”

**Natural idea #1**: estimate the support of the distribution, then we’re back in business.

Too expensive (near-linear in support size [VV11]).

**Natural idea #2**: look at *moments* (collisions).

\[ \|p\|_2^2 = \sum_i p(i)^2, \|p\|_3^3 = \sum_i p(i)^3 \]

It’s all symmetric anyway.
Upper bound: idea

**Lemma (Easy).** If $p$ is a uniform distribution,

$$||p||_3^3 = ||p||_2^4$$

**Lemma (Key).** If $p$ is $\varepsilon$-far from any uniform distribution,

$$||p||_3^3 > (1+\Phi(\varepsilon))||p||_2^4$$

**Algorithm.**
- Take samples until you see enough 2-wise collisions. (Estimate $||p||_2$)
- Take samples until you see enough 3-wise collisions. (Estimate $||p||_3$)
- Stop taking samples, and check the relation between $||p||_2^4$ and $||p||_3^3$
Upper bound

**Theorem.** There exists an (efficient) *adaptive* tester for generalized uniformity testing with *expected* sample complexity $O(1/(\|p\|_3 \varepsilon^6))$.

So... “$O(n^{2/3})$.”
Is that tight?

Yes. (For constant \( \varepsilon \).)
Theorem. For any fixed non-uniform distribution q, distinguishing between (i) \( p \equiv q \) and (ii) \( p \) uniform requires \( \Omega(1/\|p\|_3) \) samples from \( p \).

Where \( p \equiv q \) means “equal up to a permutation/relabeling.”

Equivalently, lower bound against testers which see the fingerprints/histograms.
Lower bound, instance-specific

- Strong, instance-specific (not worst-case)

- No dependence on $\varepsilon := TV(q, \emptyset)$ (sadly)

- Proven by using the framework of Paul Valiant [Valiant11]: moment-matching and **Wishful Thinking Theorem**.
Remarks and Future Directions

- **Follow-up work** of Diakonikolas, Kane, and Stewart’17: (i) improve upper bound; (ii) complement it with (worst-case) matching lower bound.

- The “right” setting for many testing questions?

- Improve dependence on $\varepsilon$

- **Instance-specific** lower bounds (a mouthful, but...)

- Lunch.
Thank you.