Testing with Alternative Distances

Gautam “G” Kamath
FOCS 2017 Workshop: Frontiers in Distribution Testing
October 14, 2017

Based on joint works with

Jayadev Acharya
Cornell

Constantinos Daskalakis
MIT

John Wright
MIT
The story so far...

- Test whether \( p = q \) versus \( \ell_1(p, q) \geq \varepsilon \)
  - Domain of \([n]\)
  - Success probability \( \geq 2/3 \)
  - Goal: Strongly sublinear sample complexity
    - \( O(n^{1-\gamma}) \) for some \( \gamma > 0 \)
- Identity testing (samples from \( p \), known \( q \))
  - \( \Theta(\sqrt{n}/\varepsilon^2) \) samples
  - [BFFKR’01, P’08, VV’14]
- Closeness testing (samples from \( p, q \))
  - \( \Theta(\max\{n^{2/3}/\varepsilon^{4/3}, \sqrt{n}/\varepsilon^2\}) \) samples
  - [BFRSW’00, V’11, CDVV’14]
Generalize: Different Distances

- \( p = q \) or \( \ell_1(p, q) \geq \varepsilon \)?

- \( d_1(p, q) \leq \varepsilon_1 \) or \( d_2(p, q) \geq \varepsilon_2 \)?
Generalize: Different Distances

- $d_1(p, q) \leq \varepsilon_1$ or $d_2(p, q) \geq \varepsilon_2$?
  - Are $p$ and $q$ $\varepsilon_1$-close in $d_1(., .)$, or $\varepsilon_2$-far in $d_2(., .)$?
  - Distances of interest: $\ell_1, \ell_2, \chi^2, KL, Hellinger$

- Classic identity testing: $\varepsilon_1 = 0, d_2 = \ell_1$

- Can we characterize sample complexity for each pair of distances?
  - Which distribution distances are sublinearly testable? [DKW’18]

- Wait, but... why?
Wait, but... why?

1. Tolerance for model misspecification
2. Useful as a proxy in classical testing problems
   • $d_1$ as $\chi^2$ distance is useful for composite hypothesis testing
   • Monotonicity, independence, etc. [ADK’15]
3. Other distances are natural in certain testing settings
   • $d_2$ as Hellinger distance is sometimes natural in multivariate settings
   • Bayes networks, Markov chains [DP’17, DDG’17]
   • Costis’ talk
Wait, but... why?

1. **Tolerance for model misspecification**

2. Useful as a proxy in classical testing problems
   - $d_1$ as $\chi^2$ distance is useful for composite hypothesis testing
   - Monotonicity, independence, etc. [ADK’15]

3. Other distances are natural in certain testing settings
   - $d_2$ as Hellinger distance is sometimes natural in multivariate settings
   - Bayes networks, Markov chains [DP’17,DDG’17]
   - Costis’ talk
Tolerance

• Is $p$ equal to $q$, or are they far from each other?
  • But why do we know $q$ exactly?

• Models are inexact
  • Measurement errors
  • Imprecisions in nature

• $p$, $q$ may be “philosophically” equal, but not literally equal

• When can we test $d_1(p, q) \leq \varepsilon_1$ versus $\ell_1(p, q) \geq \varepsilon$?
Tolerance

- \( d_1(p, q) \leq \varepsilon_1 \) vs. \( \ell_1(p, q) \geq \varepsilon \)?
  - What \( d_1 \)? How about \( \ell_1 \)?
- \( \ell_1(p, q) \leq \varepsilon / 2 \) vs. \( \ell_1(p, q) \geq \varepsilon \)?
  - No! \( \Theta(n / \log n) \) samples [VV’10]
- Chill out, relax...
- \( \chi^2 \)-distance: \( \chi^2(p, q) = \sum_{i \in \Sigma} \frac{(p_i - q_i)^2}{q_i} \)
  - Cauchy-Schwarz: \( \chi^2(p, q) \geq \ell_1^2(p, q) \)
- \( \chi^2(p, q) \leq \varepsilon^2 / 4 \) vs. \( \ell_1(p, q) \geq \varepsilon \)?
  - Yes! \( O(\sqrt{n}/\varepsilon^2) \) samples [ADK’15]
Details for a $\chi^2$-Tolerant Tester

• Goal: Distinguish (i) $\chi^2(p, q) \leq \frac{\varepsilon^2}{2}$ versus (ii) $\ell_1(p, q) \geq \varepsilon^2$

• Draw $\text{Poisson}(m)$ samples from $p$ ("Poissonization")
  • $N_i$: number of appearances of symbol $i$
  • $N_i \sim \text{Poisson}(m \cdot p_i)$
  • $N_i$’s are now independent!

• Statistic: $Z = \sum_{i \in \Sigma} \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
Details for a $\chi^2$-Tolerant Tester

- Goal: Distinguish (i) $\chi^2(p, q) \leq \frac{\varepsilon^2}{2}$ versus (ii) $\ell_1^2(p, q) \geq \varepsilon^2$

- Statistic: $Z = \sum_{i \in [n]} \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
  - $N_i$: # of appearances of $i$; $m$: # of samples
  - $E[Z] = m \cdot \chi^2(p, q)$
    - (i): $E[Z] \leq m \cdot \frac{\varepsilon^2}{2}$, (ii): $E[Z] \geq m \cdot \varepsilon^2$
  - Can bound variance of $Z$ with some work
    - Need to avoid low prob. elements of $q$
      1. Either ignore $i$ such that $q_i \leq \frac{\varepsilon^2}{10n}$; or
      2. Mix lightly ($O(\varepsilon^2)$) with uniform distribution (also in [G’16])
  - Apply Chebyshev’s inequality

Side-Note:
- Pearson’s $\chi^2$-test uses statistic $\sum_{i} \frac{(N_i - m \cdot q_i)^2}{m \cdot q_i}$
- Subtracting $N_i$ in the numerator gives an unbiased estimator and importantly may hugely decrease variance
  - [Zelterman’87]
  - [VV’14, CDVV’14, DKN’15]
# Tolerant Identity Testing

<table>
<thead>
<tr>
<th>Distribution Distance</th>
<th>Testability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = q$</td>
<td>$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Pan08]</td>
</tr>
<tr>
<td>$d_{\chi^2}(p, q) \leq \varepsilon^2/4$</td>
<td>$O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 1]</td>
</tr>
<tr>
<td>$d_{KL}(p, q) \leq \varepsilon^2/4$</td>
<td>$\Omega\left(\frac{n}{\log n}\right)$ [Theorem 8]</td>
</tr>
<tr>
<td>$d_H(p, q) \leq \varepsilon/2\sqrt{2}$</td>
<td>$O\left(\frac{n}{\log n}\right)$ [Corollary 3]</td>
</tr>
<tr>
<td>$d_{TV}(p, q) \leq \varepsilon/2$ or $\varepsilon^2/4$</td>
<td>$O\left(\frac{n}{\log n}\right)$ [Corollary 3]</td>
</tr>
</tbody>
</table>

$\ell_2$-distance versus TV-distance:

$\ell_2(p, q) \leq \frac{\varepsilon}{\sqrt{n}}$ vs $d_{TV}(p, q) \geq \varepsilon$

$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 2]

---

(Daskalakis, K., Wright. Which Distribution Distances are Sublinearly Testable? SODA 2018.)
Tolerant Testing Takeaways

1. Can handle $\ell_2$ or $\chi^2$ tolerance at no additional cost
   - $\Theta(\sqrt{n}/\epsilon^2)$ samples

2. KL, Hellinger, or $\ell_1$ tolerance are expensive
   - $\Theta(n/\log n)$ samples
   - KL result based off hardness of entropy estimation

3. Closeness testing ($q$ unknown): Even $\chi^2$ tolerance is costly!
   - $\Theta(n/\log n)$ samples
   - Only $\ell_2$ tolerance is free
   - Proven via hardness of $\ell_1$-tolerant identity testing
   - Since $q$ is unknown, $\chi^2$ is no longer a polynomial
Application: Testing for Structure

• Composite hypothesis testing
• Test against a class of distributions!
  • \( p \in \mathcal{C} \) versus \( \ell_1(p, \mathcal{C}) > \varepsilon \)
    \[
    \min_{q \in \mathcal{C}} \ell_1(p, q)
    \]
• Example: \( \mathcal{C} = \) all monotone distributions
  • \( p_i \)'s are monotone non-increasing
  • Others: unimodality, log-concavity, monotone hazard rate, independence
  • All can be tested in \( \Theta(\sqrt{n}/\varepsilon^2) \) samples [ADK'15]
    • Same complexity as vanilla uniformity testing!
Testing by Learning

• Goal: Distinguish $p \in C$ from $\ell_1(p, C) > \epsilon$

• Learn-then-Test:
  1. Learn hypothesis $q \in C$ such that
     • $p \in C \Rightarrow \chi^2(p, q) \leq \epsilon^2 / 2$ (needs cheap “proper learner” in $\chi^2$)
     • $\ell_1(p, C) > \epsilon \Rightarrow \ell_1(p, q) > \epsilon$ (automatic since $q \in C$)
  2. Perform “tolerant testing”
     • Given sample access to $p$ and description of $q$, distinguish $\chi^2(p, q) \leq \epsilon^2 / 2$ from $\ell_1(p, q) > \epsilon$

• Tolerant testing (step 2) is $O(\sqrt{n} / \epsilon^2)$
  • Naïve approach (using $\ell_1$ instead of $\chi^2$) would require $\Omega(n / \log n)$

• Proper learners in $\chi^2$ (step 1)?
  • Claim: This is cheap
Hellinger Testing

- Change $d_2$ instead of $d_1$

- Hellinger distance: $H^2(p, q) = \frac{1}{2} \sum_{i \in [n]} (\sqrt{p_i} - \sqrt{q_i})^2$
  - Between linear and quadratic relationship with $\ell_1$
  - $H^2(p, q) \leq \ell_1(p, q) \leq H(p, q)$

- Natural distance when considering a collection of iid samples
  - Comes up in some multivariate testing problems (Costis @ 2:55)

- Testing $p = q$ vs. $H(p, q) \geq \varepsilon$?

- Trivial results via $\ell_1$ testing
  - Identity: $O(\sqrt{n}/\varepsilon^4)$ samples
  - Closeness: $O(\max\{n^{2/3}/\varepsilon^{8/3}, \sqrt{n}/\varepsilon^4\})$ samples
Hellinger Testing

- Testing $p = q$ vs. $H(p, q) \geq \epsilon$?
- Trivial results via $\ell_1$ testing
  - Identity: $O(\sqrt{n}/\epsilon^4)$ samples
  - Closeness: $O(\max\{n^{2/3}/\epsilon^{8/3}, \sqrt{n}/\epsilon^4\})$ samples
- But you can do better!
  - Identity: $\Theta(\sqrt{n}/\epsilon^2)$ samples
    - No extra cost for $\ell_2$ or $\chi^2$ tolerance either!
  - Closeness: $\Theta(\min\{n^{2/3}/\epsilon^{8/3}, n^{3/4}/\epsilon^2\})$ samples
    - LB and previous UB in [DK’16]
- Similar chi-squared statistics as [ADK’15] and [CDVV’14]
  - Some tweaks and more careful analysis to handle Hellinger

_Daskalakis, K., Wright. Which Distribution Distances are Sublinearly Testable? SODA 2018._
Miscellanea

• $p = q$ vs. $KL(p, q) \geq \varepsilon$?
  • Trivially impossible, due to ratio between $p_i$ and $q_i$
  • $p_i = \delta, q_i = 0, \delta \to 0$

• Upper bounds for $\Omega(n / \log n)$ testing problems?
  • i.e., $KL(p, q) \leq \varepsilon^2 / 4$ vs. $\ell_1(p, q) \geq \varepsilon$?
  • Use estimators mentioned in Jiantao’s talk
Thanks!

Table 1: Identity Testing

<table>
<thead>
<tr>
<th>$p = q$</th>
<th>$d_{TV}(p,q) \geq \epsilon$</th>
<th>$d_{H}(p,q) \geq \epsilon/\sqrt{2}$</th>
<th>$d_{KL}(p,q) \geq \epsilon^2$</th>
<th>$d_{\chi^2}(p,q) \geq \epsilon^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{*2}(p,q) \leq \epsilon^2/4$</td>
<td>$O\left(\frac{n^{\frac{3}{4}}}{\epsilon^{2/3}}\right)$ [Pan08]</td>
<td>$O\left(\frac{n^{\frac{1}{2}}}{\epsilon^{2/3}}\right)$ [Theorem 1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{KL}(p,q) \leq \epsilon^2/4$</td>
<td>$\Omega\left(\frac{n}{\log n}\right)$ [Theorem 8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{H}(p,q) \leq \epsilon/2\sqrt{2}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{TV}(p,q) \leq \epsilon/2$ or $\epsilon^2/4$</td>
<td>$O\left(\frac{n}{\log n}\right)$ [Corollary 3]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Equivalence Testing

<table>
<thead>
<tr>
<th>$p = q$</th>
<th>$d_{TV}(p,q) \geq \epsilon$</th>
<th>$d_{H}(p,q) \geq \epsilon/\sqrt{2}$</th>
<th>$d_{KL}(p,q) \geq \epsilon^2$</th>
<th>$d_{\chi^2}(p,q) \geq \epsilon^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{*2}(p,q) \leq \epsilon^2/4$</td>
<td>$O\left(\max\left{n^{1/3},\frac{n^{1/2}}{\epsilon^{2/3}}\right}\right)$ [CDVV14]</td>
<td>$O\left(\min\left{n^{1/4},\frac{n^{2/3}}{\epsilon^{2/3}}\right}\right)$ [Theorem 5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{KL}(p,q) \leq \epsilon^2/4$</td>
<td>$\Omega\left(\frac{n}{\log n}\right)$ [Theorem 9]</td>
<td>$\Omega\left(\min\left{n^{3/4},\frac{n^{7/6}}{\epsilon^{2/3}}\right}\right)$ [DK16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{H}(p,q) \leq \epsilon/2\sqrt{2}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d_{TV}(p,q) \leq \epsilon/2$ or $\epsilon^2/4$</td>
<td>$O\left(\frac{n}{\log n}\right)$ [Corollary 3]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>