\( \ell_p \) Testing and Learning of Discrete Distributions

Bo Waggoner
Harvard

*Thanks: Clément Canonne*

ITCS 2015
Drawing Conclusions from Data

Given i.i.d. samples from a discrete distribution $A$, what can you tell me about $A$?

This paper:
- **Learning**: Estimate $A$ “accurately”
- **Uniformity Testing**: Is $A$ uniform or “far from” uniform?
Previously studied: $\ell_1$ distance

(equivalently: total variation distance):

$$\| A - B \|_1 = \sum_{i=1}^{n} |A_i - B_i|$$
This work: $\ell_p$ distance, $p \geq 1$

$$\|A - B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{1/p}$$

$$\|A - B\|_\infty = \max_{i=1 \ldots n} |A_i - B_i|$$
This work: $\ell_p$ distance, $p \geq 1$

$$\| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}}$$

$$\| A - B \|_\infty = \max_{i=1 \ldots n} |A_i - B_i|$$

Given $n$, $\epsilon$:

**Learning:** Output $\hat{A}$ such that $\| \hat{A} - A \|_p \leq \epsilon$.

**Uniformity testing:** If $A = U$, output “unif”; if $\| A - U \|_p \geq \epsilon$, “not”.

Both cases: Except with constant failure probability $\delta$ (e.g. 1/3)
Results

![Image showing a question about how many samples are needed.]

- Upper and lower bounds for each $\ell_p$ metric.
- Matching up to constant factors in most cases.

**Unlike $\ell_1$ case:**

- Exists a sufficient # of samples independent of $n$
- Behavior differs in “small” and “large” $n$ regimes

\[ \| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]
Why care about $\ell_p$?

$\|A - B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}}$

Why Bo cares:

- I like the math/probability involved
- Fundamental problems deserve elegant algorithms/proofs (and small constants)
Why care about $\ell_p$?

Why else you might care:

- **Small data in a big world.**
  What if we do not have enough samples to draw confident $\ell_1$ conclusions?

- $\ell_p$ testers/learners are often useful as subroutines (Batu et al 2013, Diakonikolas et al 2015, ...)

\[ \|A - B\|_p = \left( \sum_{i=1}^n |A_i - B_i|^p \right)^{\frac{1}{p}} \]
What was known?

\[ \| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]

- **Learning**: order-optimal \( \ell_1 \) (folklore), also \( \ell_2 \) and \( \ell_\infty \).

- **Uniformity testing**: 
  - \( \ell_1 \): order-optimal lower, and upper for “very big” \( n \) (Paninski 2008)
  - Independently (Diakonikolas, Kane, Nikishkin 2015): order-optimal \( \ell_1 \), and \( \ell_2 \) for small-\( n \) regime

- **Note**: many cases “immediate” from prior work, most (all?) cases probably “easy” to experts

- But hopefully when taken together, **big picture insights** emerge
Outline

• Introductory stuff ✓

• Learning

• Uniformity testing

• Summary
Learning

\[ \| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]
\[ \|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]
Learning

\[ \|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{1/p} \]

For \( p > 1 \):

- Exists a sufficient \# of samples independent of \( n \)
- Behavior differs in “small” and “large” \( n \) regimes
Learning Alg

1. Let $\Pr[i] \propto \# \text{samples of } i$

$$\|A - B\|_p = \left(\sum_{i=1}^n |A_i - B_i|^p\right)^{\frac{1}{p}}$$
Learning Alg

1. Let $\Pr[i] \propto \# \text{ samples of } i$

Analysis:
- Elegant “folklore” proof for $\ell_2$ (thanks Clément!)
- Clément and I extended to general $\ell_p$ and large-$n$ cases

Theorem (in particular):
- For $p = 1$, $\frac{1}{\delta} \frac{n}{\epsilon^2}$ samples are sufficient to learn.
- For $p \geq 2$, $\frac{1}{\delta} \frac{1}{\epsilon^2}$ samples are sufficient to learn.
Learning Alg

1. Let $Pr[i]$ = # samples of $i$

Analysis:

- Elegant "folklore" proof for $\ell_2$ (thanks Clément!)
- Tweaks for $\ell_p$ and large-$n$ cases

Theorem (in particular):

- For $p = 1$, samples are sufficient to learn.
- For $p \geq 2$, samples are sufficient to learn.

Given $p$, consider Holder conjugate $q : \frac{1}{p} + \frac{1}{q} = 1$

\[ \|A - B\|_p = \left( \sum_{i=1}^n |A_i - B_i|^p \right)^{\frac{1}{p}} \]

\[
\begin{align*}
  p: & \quad 1 \quad \frac{5}{4} \quad \frac{4}{3} \quad \frac{3}{2} \quad 2 \quad \ldots \quad \infty \\
  q: & \quad \infty \quad 5 \quad 4 \quad 3 \quad 2 \quad \ldots \quad 1
\end{align*}
\]

- For $p \geq 2$, $\frac{\delta}{\epsilon^2}$ samples are sufficient to learn.

small-$n$ regime: $n \leq \frac{1}{\epsilon^q}$

large-$n$ regime: $n \geq \frac{1}{\epsilon^q}$
Learning

For $p > 1$:

- Exists a sufficient # of samples independent of $n$
- Behavior differs in “small” and “large” $n$ regimes

**Threshold:** $n = \frac{1}{\epsilon^q}$

$$\|A - B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}}$$
Outline

• Introductory stuff ✓

• Learning ✓

• Uniformity testing

• Summary
Classic Coin Question

Coin: either fair or one side with $\epsilon$ more probability.

Q: How many flips to tell?

A: $O\left(\frac{1}{\epsilon^2}\right)$. 
6-sided die: either fair or one side with $\epsilon$ more probability.

Q: Do we need more trials than the coin, or fewer?
Classic Dice Question?

6-sided die: either fair or one side with $\epsilon$ more probability.

Q: Do we need more trials than the coin, or fewer?
A: Fewer!
Classic Dice Question?

6-sided die: either fair or one side with $\epsilon$ more probability.

Q: Do we need more trials than the coin, or fewer?

A: Fewer!
Classic Dice Question?

6-sided die: either fair or one side with $\epsilon$ more probability.

Q: Do we need more trials than the coin, or fewer?

A: Fewer! ($\ell_\infty$)

For $\ell_1$, need more.
In between?
Testing, $1 \leq p \leq 2$

$$\| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}}$$
Testing Alg

\[ \| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{1/p} \]

**Collision:** pair of samples that are both of the same coordinate

Testing Alg

\[ \|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{1/p} \]

1. Let \( C = \# \) collisions

2. Pick threshold \( T \)

3. If \( C \leq T \), output “uniform”; else, “not”.

Alg is optimal for all \( 1 \leq p \leq 2 \), all regimes! (by selecting \( \# \) samples and \( T \) appropriately)
Testing Alg

\[ \|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]

1. Let \( C = \# \) collisions

2. Pick threshold \( T \)

3. If \( C \leq T \), output “uniform”; else, “not”.

Alg is optimal for all \( 1 \leq p \leq 2 \), all regimes! (by selecting \# samples and \( T \) appropriately)

**Theorem (in particular):**

- For \( p = 1 \), \( \frac{9 \sqrt{n}}{\delta \epsilon^2} \) samples are sufficient to test uniformity.

- For \( p = 2 \), \( \max \left\{ \frac{9}{\delta \sqrt{n} \epsilon^2}, \frac{9 \frac{1}{\epsilon}}{\delta} \frac{1}{\epsilon} \right\} \) samples suffice.
Testing, $1 \leq p \leq 2$

Let $A - B \in \mathbb{R}^{n \times n}$ be a matrix. The $p$-norm of the difference $A - B$ is defined as

$$\|A - B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{1/p}$$

Threshold: $n = \frac{1}{\epsilon^q}$
$\ell_\infty$ Testing

$$\|A-B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}}$$
\[ \ell_\infty \text{ Testing} \]

\[ \|A - B\|_p = \left( \sum_{i=1}^n |A_i - B_i|^p \right)^{\frac{1}{p}} \]

Theorem (for \( p = \infty \)):
- If \( \Theta \left( \frac{n}{\log n} \right) \leq \frac{1}{\epsilon} \) ("small"), \( \Theta \left( \frac{\log n}{n \epsilon^2} \right) \) samples are necessary/sufficient.
- If \( \Theta \left( \frac{n}{\log n} \right) \geq \frac{1}{\epsilon} \) ("large"), \( \Theta \left( \frac{1}{\epsilon} \right) \) samples are necessary/sufficient.

Note:
- Still have "small" and "large" regimes, but \( \log(n) \) gets involved (Bounds still match at threshold)
\[ \|
A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]

**Theorem (for \( p = \infty \)):**

- If \( \Theta\left( \frac{n}{\log n} \right) \leq \frac{1}{\epsilon} \) ("small"), \( \Theta\left( \frac{\log n}{n\epsilon^2} \right) \) samples are necessary/sufficient.

- If \( \Theta\left( \frac{n}{\log n} \right) \geq \frac{1}{\epsilon} \) ("large"), \( \Theta\left( \frac{1}{\epsilon} \right) \) samples are necessary/sufficient.

**Note:**
- Still have "small" and "large" regimes, but \( \log(n) \) gets involved (Bounds still match at threshold)

**Alg:**
- Small-\( n \): look for "outlier" coordinate
- Large-\( n \): “bucket” into \( n^* \) groups and look for outlier bucket
Gap for $2 < p < \infty$

- $\ell_2$ alg $\rightarrow$ sufficient
  $\ell_\infty$ bound $\rightarrow$ necessary

- Gap only in small-$n$ case

- Seems to need different ideas

\[ \| A - B \|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]
Outline

- Introductory stuff ✓
- Learning ✓
- Uniformity testing ✓
- Summary
Algorithms Summary

- **Learning**: naive alg is order-optimal everywhere
- **Uniformity testing**: Collision Tester is order-optimal for $1 \leq p \leq 2$
- **Uniformity testing for $\ell_\infty$**: “almost-naive” alg is order-optimal
Ideas Summary

For $p > 1$:

- Exists a sufficient # of samples independent of $n$
- Behavior differs in “small” and “large” $n$ regimes
- $\frac{1}{\epsilon^q}$ seems to upper-bound “apparent support size”
Future Work

- Close gap for uniformity testing, $2 < p < \infty$, small $n$
- Strengthen “tightness” of lower bound for small-$n$ learning, $1 \leq p < 2$

- Test and learn “thin” distributions?
- Test and learn when $n$ is not known?
- Test and learn for other “exotic” metrics? (Do Ba, Nguyen, Nguyen, Rubinfeld 2011)

$$\|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}}$$
Future Work

\[ \|A - B\|_p = \left( \sum_{i=1}^{n} |A_i - B_i|^p \right)^{\frac{1}{p}} \]

- Close gap for uniformity testing, \(2 < p < \infty\), small \(n\)
- Strengthen “tightness” of lower bound for small-\(n\) learning, \(1 \leq p < 2\)
- Test and learn “thin” distributions?
- Test and learn when \(n\) is not known?
- Test and learn for other “exotic” metrics? (Do Ba, Nguyen, Nguyen, Rubinfeld 2011)

Thanks!