Clément gave me a lot of help, ideas, advice. We first started talking about the problem due to a cstheory.stackexchange.com post.
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 paper considers the same questions for general $l_p$ metrics. The functional form isn’t important, main point is that:
- defined for all real $p \geq 1$
- $l1$ is Manhattan distance
- $l2$ is Euclidean distance
- as we increase $p$, we put more emphasis on few “heavy” elements
- extreme case is $l\infty$ which only measures largest difference
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

\[ \|A - B\|_p = \left( \sum_{j=1}^{n} |A_j - B_j|^p \right)^{1/p} \]

• 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
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 $\ell_p$?  

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

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, ...)

---

It will turn out that we can often draw $\ell_p$ conclusions with far fewer samples, especially over large distributions.
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 $\ell_\infty$.
  \[ O\left( \frac{n}{\epsilon^2} \right) \]

- **Uniformity testing**:
  \[ O\left( \frac{\sqrt{n}}{\epsilon^2} \right) \]
  - $\ell_2$: 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

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

Emperor's new plot

Think of the epsilon tolerance as 0.01 or something. Now we'll think about support size $n$ in terms of powers of $1/\epsilon$. The question is how many samples we need as $n$ changes. Note the plot is in log-log scale.
Learning

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

Starting point: known bounds look like this.
Learning

For $p > 1$:

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

Here's what bounds look like for learning, necessary and sufficient up to constant factors, for 5 particular choices of $l_p$ metric. Note $l_p$ for $2 \leq p \leq \infty$ is always $1/\epsilon^2$ samples.

In between 1 and 2, we have a small-n regime where the sample complexity increases, then a large-n regime where it's constant. Before we see what the bounds are, let's see the algorithm.
Learning Alg

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

1. Let \( \Pr[i] \propto \# \text{ samples of } i \)
Learning Alg

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

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.

There's no big-O in the theorem – the constant is 1!
Learning Alg

\[ ||A - B||_p = \left( \sum_{j=1}^{n} |A_j - B_j|^p \right)^{\frac{1}{p}} \]

1. Let \( D(x) \) = # samples of \( x \)

Analysis
- El
- TV

Given \( p \), consider Holder conjugate \( q \):
\[
\frac{1}{p} + \frac{1}{q} = 1
\]

<table>
<thead>
<tr>
<th>( p )</th>
<th>1</th>
<th>5/4</th>
<th>4/3</th>
<th>3/2</th>
<th>2</th>
<th>( \ldots )</th>
<th>( \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td>( \infty )</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>( \ldots )</td>
<td>1</td>
</tr>
</tbody>
</table>

- For \( p \geq 2 \), \( \frac{\|x\|_p}{\epsilon^p} \) samples are sufficient to learn.
- For \( p > 1 \), a key threshold is \( \frac{1}{\epsilon^q} \).

It turns out the conjugate pairs, as in analysis, become important.
For \( p > 1 \), a key threshold is \( 1/\epsilon^q \).
For $p > 1$:
- Exists a sufficient # of samples independent of $n$
- Behavior differs in “small” and “large” $n$ regimes

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

In general, for the small-$n$ regime we have the bound shown (exact form not important for this talk), and for the large-$n$ regime the bound is $1/\epsilon^q$, which is interesting because the “threshold” for large-$n$ is $1/\epsilon^q$. 
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: $o\left(\frac{1}{\epsilon^2}\right)$. 
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?

I don't know of anyone who asked this question before.
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!

Intuition: With 2-sided coin, large variance in the counts of heads and tails. Need more flips for the bias to “overwhelm” the variance.

With 6-sided die, each side has smaller variance.
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?

That was an $l$-infinity question since we had one outlier coordinate.
On the other hand, for $l$-1 problems we need more samples.
Testing, $1 \leq p \leq 2$

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

For lp uniformity testing with $p=4/3$, for every support size $n$, theta($1/\epsilon^2$) samples is necessary and sufficient (whether you have a coin, or a die, or a lottery, or whatever). For $p < 4/3$, increasing in $n$ in small-$n$ regime, then constant. For $p > 4/3$, decreasing then constant.
Testing Alg

\[ \|A - B\|_p = \left( \sum_{j=1}^{n} |A_j - B_j|^p \right)^{\frac{1}{p}} \]

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


Not the expected number of collisions when drawing m samples from A is

\[
\binom{m}{2} \|A\|_2^2 \]

\[
= \binom{m}{2} \left( \|U\|_2^2 + \|A-U\|_2^2 \right) 
\]

\[
= \binom{m}{2} \left( \frac{1}{n} + \|A-U\|_2^2 \right). 
\]

So the L2 distance to uniformity directly controls the expected number of samples.
Testing Alg

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

1. Let \( C = \# \text{ collisions} \)
2. Pick threshold \( T \)
3. If \( C \leq T \), output \( \text{“uniform”}; \) else, \( \text{“not”}. \)

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

Point: uniform distribution minimizes number of collisions.
Testing Alg

\[ ||A - B||_p = \left( \sum_{j=1}^{n} |A_j - B_j|^p \right)^{\frac{1}{p}} \]

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

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

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

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

For small-$n$ regime, bound isn't so important.
For large-$n$ regime, it is $\sqrt[4]{1/\epsilon^q}$, interesting because $n = 1/\epsilon^q$ is the threshold.
The blue line is the sample complexity for l2 testing; green is linfinity. So it decreases more sharply and is then constant at 1/epsilon.
\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"), \( \Theta \left( \frac{\log n}{n^{\epsilon}} \right) \) 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)

![Graph showing bounds and sample sizes](image)

Actually I'm quite happy to have worked this out cleanly (tight everywhere to constant factors). Note that at the threshold between large and small \( n \), the bounds match.
Here, $n^*$ is the “threshold” $n$, the value where $\Theta(n^*/\log(n^*)) = 1/\epsilon$. So when $n$ is large, no matter how large it is, group the coordinates into $n^*$ groups and pretend it's the uniform distribution on support $n^*$.

The proof here is just chernoff bound on each coordinate (or bucket) and union-bound over the coordinates (buckets). The cool thing is it's tight to constant factors.
Gap for $2 < p < \infty$

\[ \|A - B\|_p = \left( \sum_{j=1}^{\infty} |A_j - B_j|^p \right)^{1/p} \]

- $\ell_2$ alg $\rightarrow$ sufficient
- $\ell_\infty$ bound $\rightarrow$ necessary
- Gap only in small-$n$ case
- Seems to need different ideas

So the blue line is an upper bound, green is a lower bound, for every $l^p$ metric with $2 < p < \text{infinity}$. 
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 $\ell_p$**: “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
- $\frac{1}{e^t}$ seems to upper-bound “apparent support size”
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)

By “thin”, I mean small \(l\)-infty norm (every coordinate has small probability). Should definitely be easier to e.g. learn thin distributions for at least some \(lp\) metrics.
Future Work

\[ \|A - B\|_p = \left| \sum_{j=1}^{n} |A_j - B_j|^p \right|^\frac{1}{p} \]

- 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!