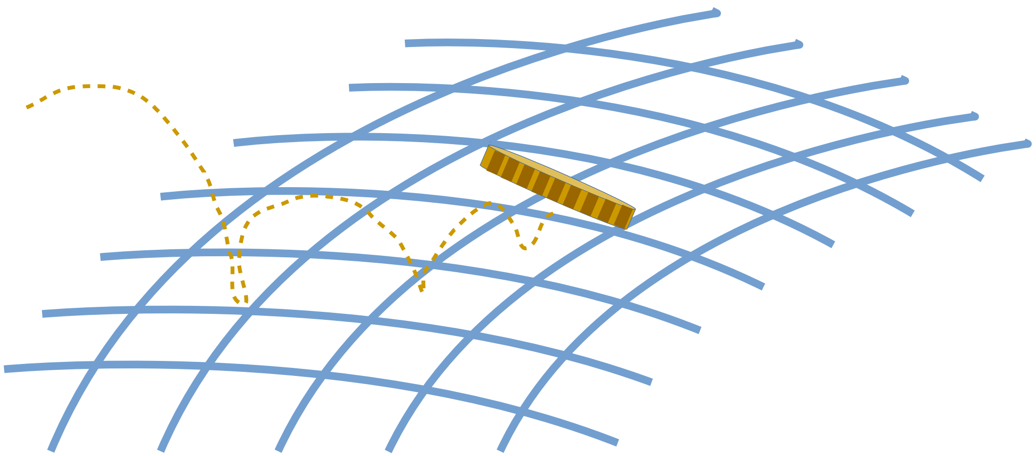
ℓ_p Testing and Learning of Discrete Distributions

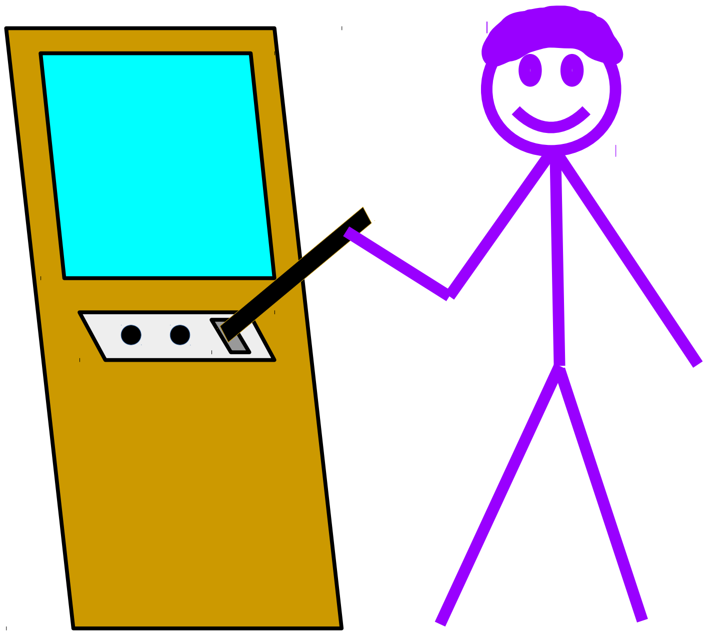


Bo Waggoner
Harvard

*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?

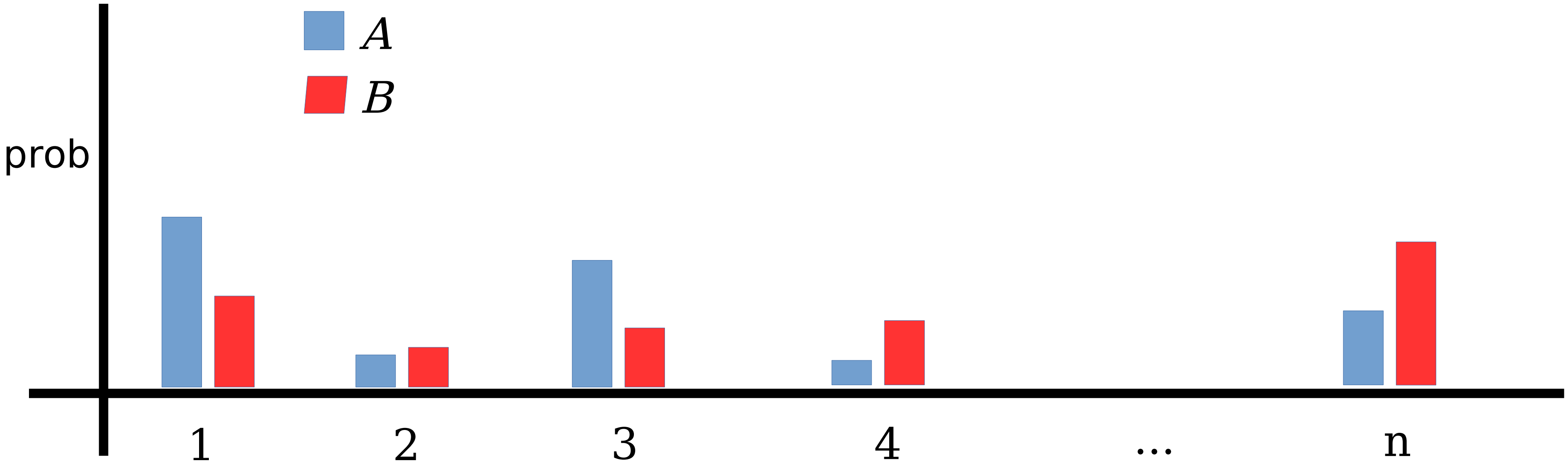
This paper:

- **Learning:** Estimate A “accurately”
- **Uniformity Testing:**
  Is A uniform or “far from” uniform?

Previously studied: ℓ_1 distance

(equivalently: total variation distance):

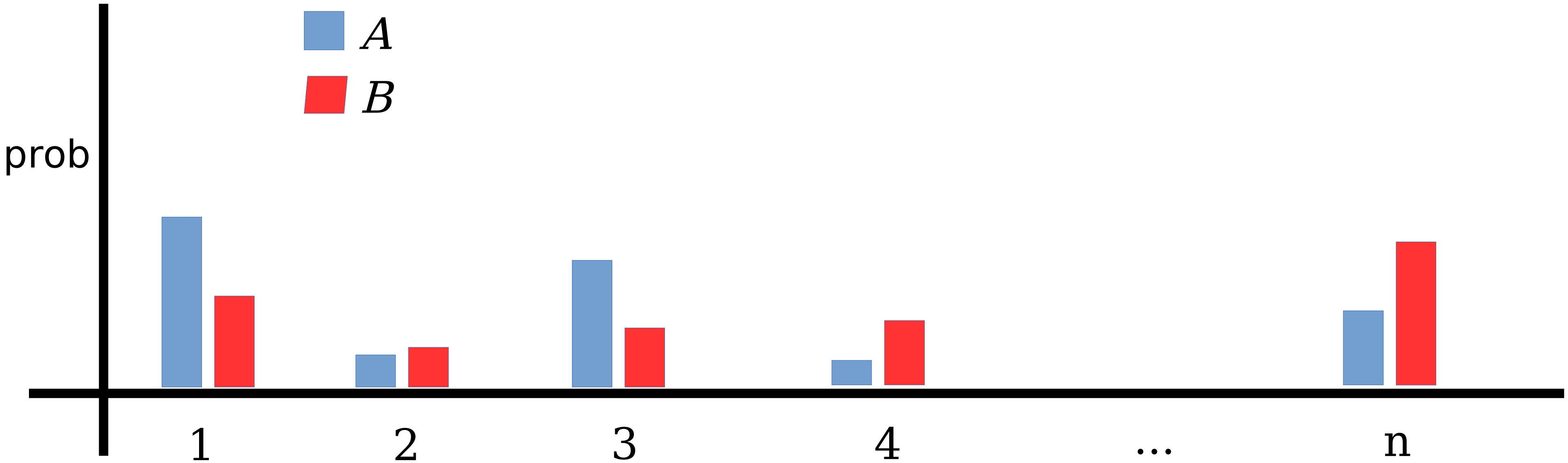
$$\|A-B\|_1 = \sum_{i=1}^{n} |A_i - B_i|$$



This work: ℓ_p distance, $p \geq 1$

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

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



This work: ℓ_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, ϵ:

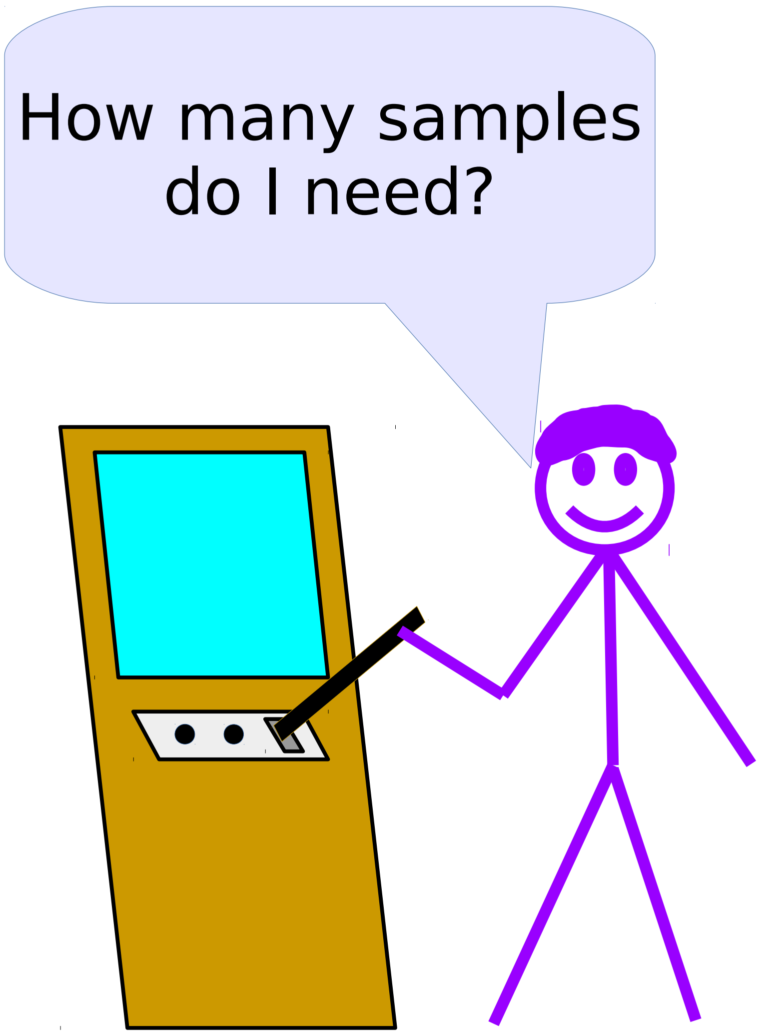
Learning: Output $\hat{A}$ such that $\|\hat{A} - A\|_p \leq \epsilon$.

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

Both cases: Except with constant failure probability δ (e.g. 1/3)

Results

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



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

Unlike ℓ_1 case:

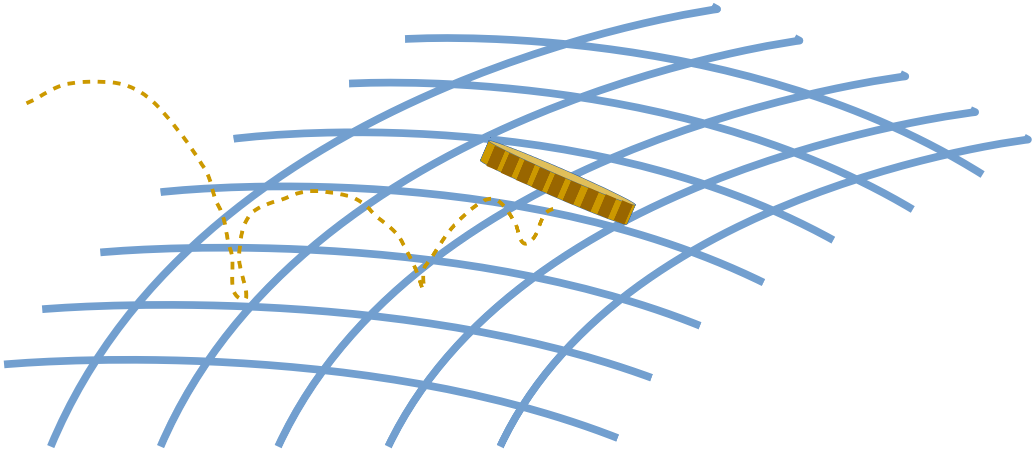
- Exists a sufficient # of samples independent of n
- Behavior differs in “small” and “large” n regimes

Why care about ℓ_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)

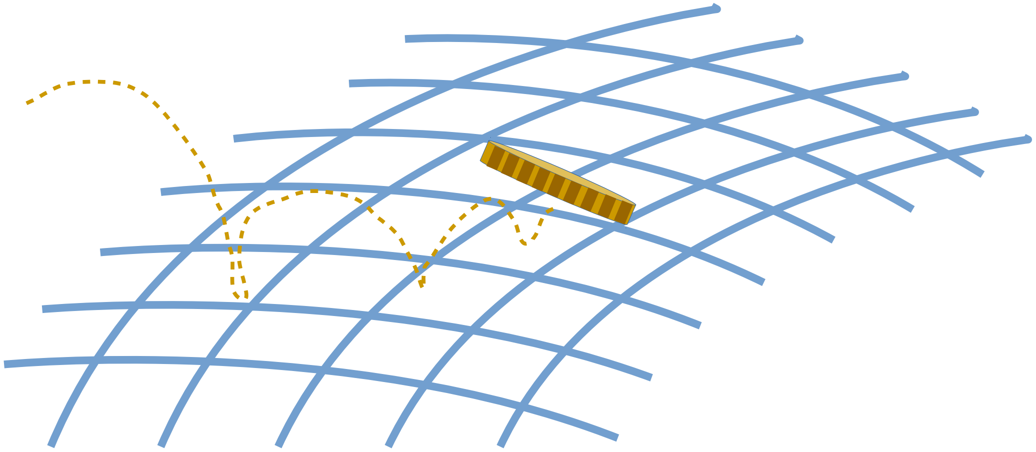


Why care about ℓ_p?

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

Why else you might care:

- ***Small data in a big world***.
  What if we do not have enough samples to draw confident ℓ_1 conclusions?
- ℓ_p testers/learners are often useful as subroutines
  (Batu et al 2013, Diakonikolas et al 2015, …)



What was known?

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

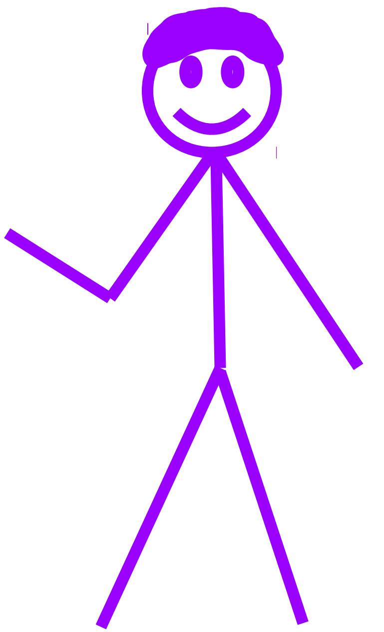
- **Learning**: order-optimal ℓ_1 (folklore), also ℓ_2 and ℓ_∞. $O\left(\frac{n}{\epsilon^2}\right)$

- **Uniformity testing**: $O\left(\frac{\sqrt{n}}{\epsilon^2}\right)$
  - ℓ_1: order-optimal lower, and upper for "very big" n (Paninski 2008)
  - Independently (Diakonikolas, Kane, Nikishkin 2015): order-optimal ℓ_1, and ℓ_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

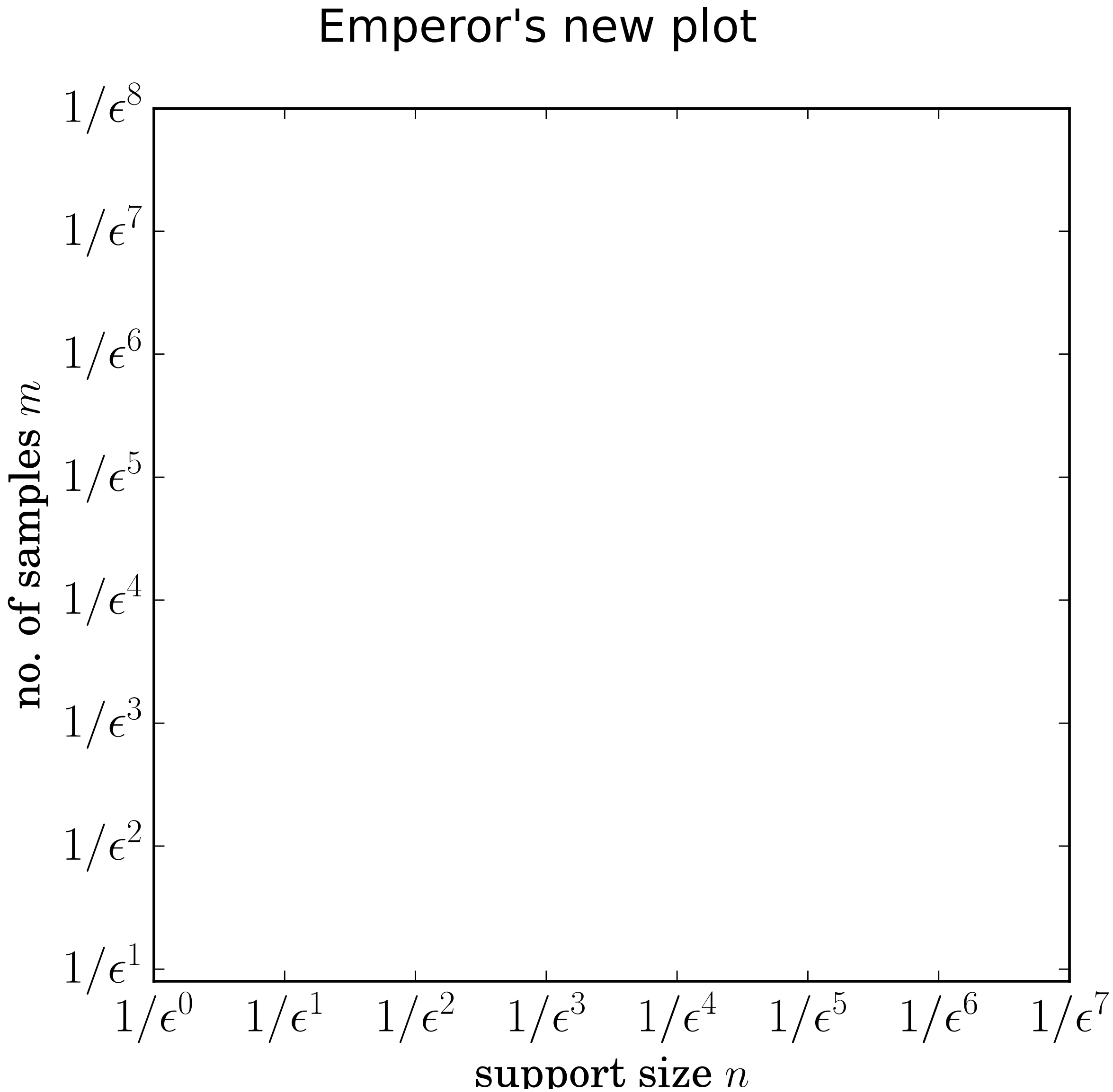
Outline

- Introductory stuff ✓
- Learning
- Uniformity testing
- Summary



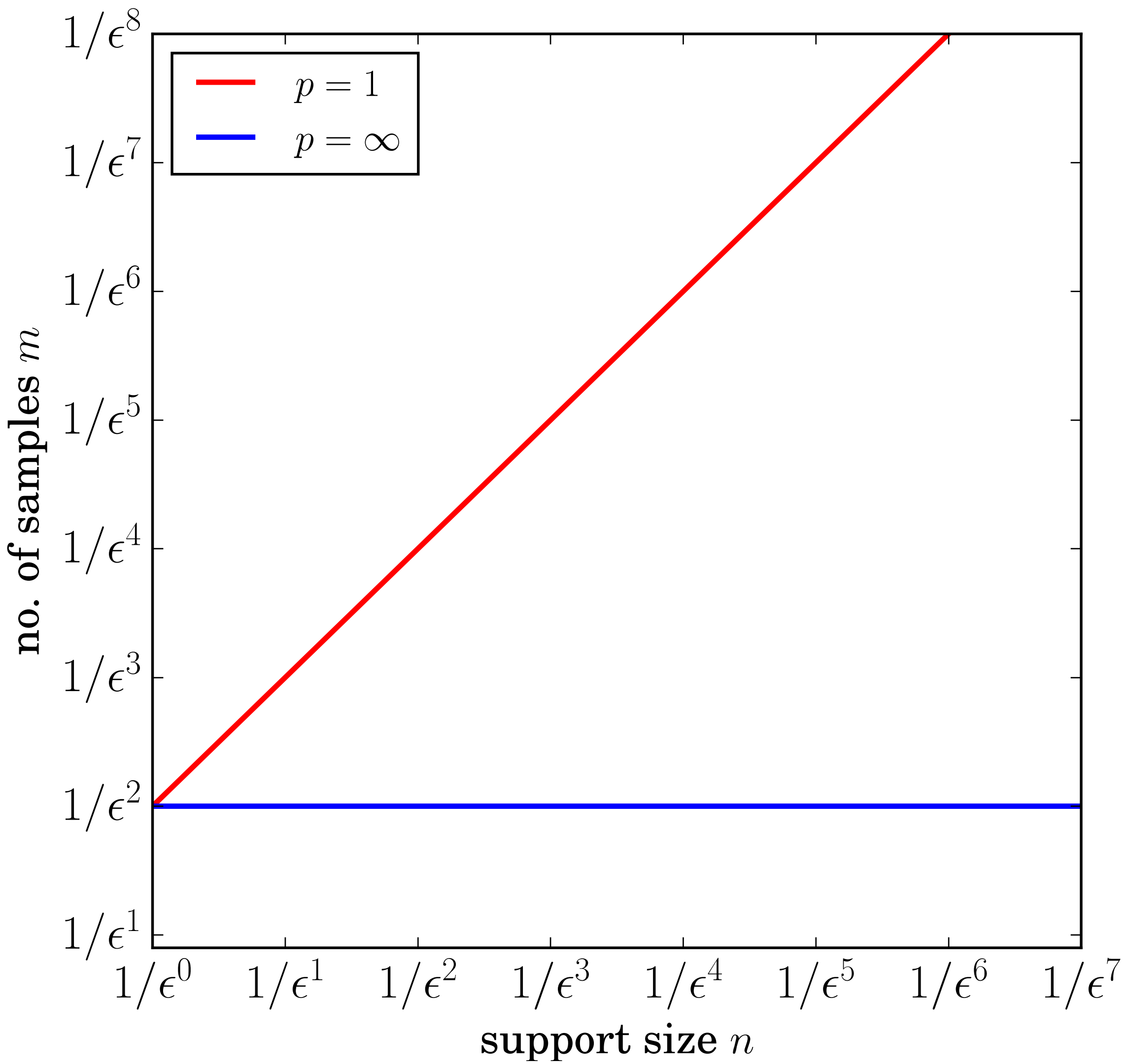
Learning

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



Learning

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

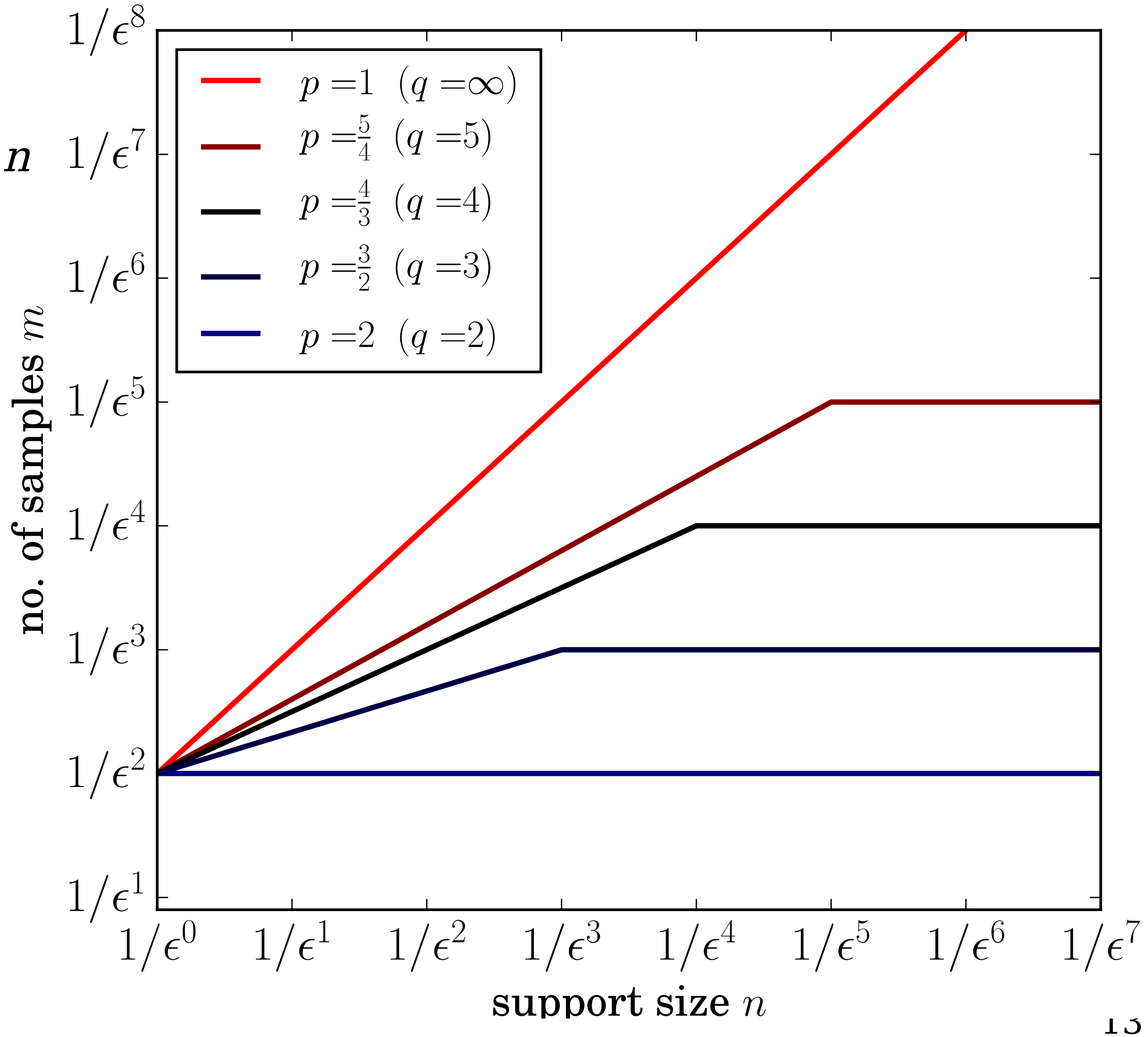


Learning

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

For $p > 1$:

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



Learning Alg

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

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

Learning Alg

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

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

Analysis:

- Elegant “folklore” proof for ℓ_2 (thanks Clément!)
- Clément and I extended to general ℓ_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

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

1. L... Pr[i] ... # samples of i

Ana...

- El...

- Tw...

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

p:	1	$\frac{5}{4}$	$\frac{4}{3}$	$\frac{3}{2}$	2	$\dots$	∞
q:	∞	5	4	3	2	$\dots$	1

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

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

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

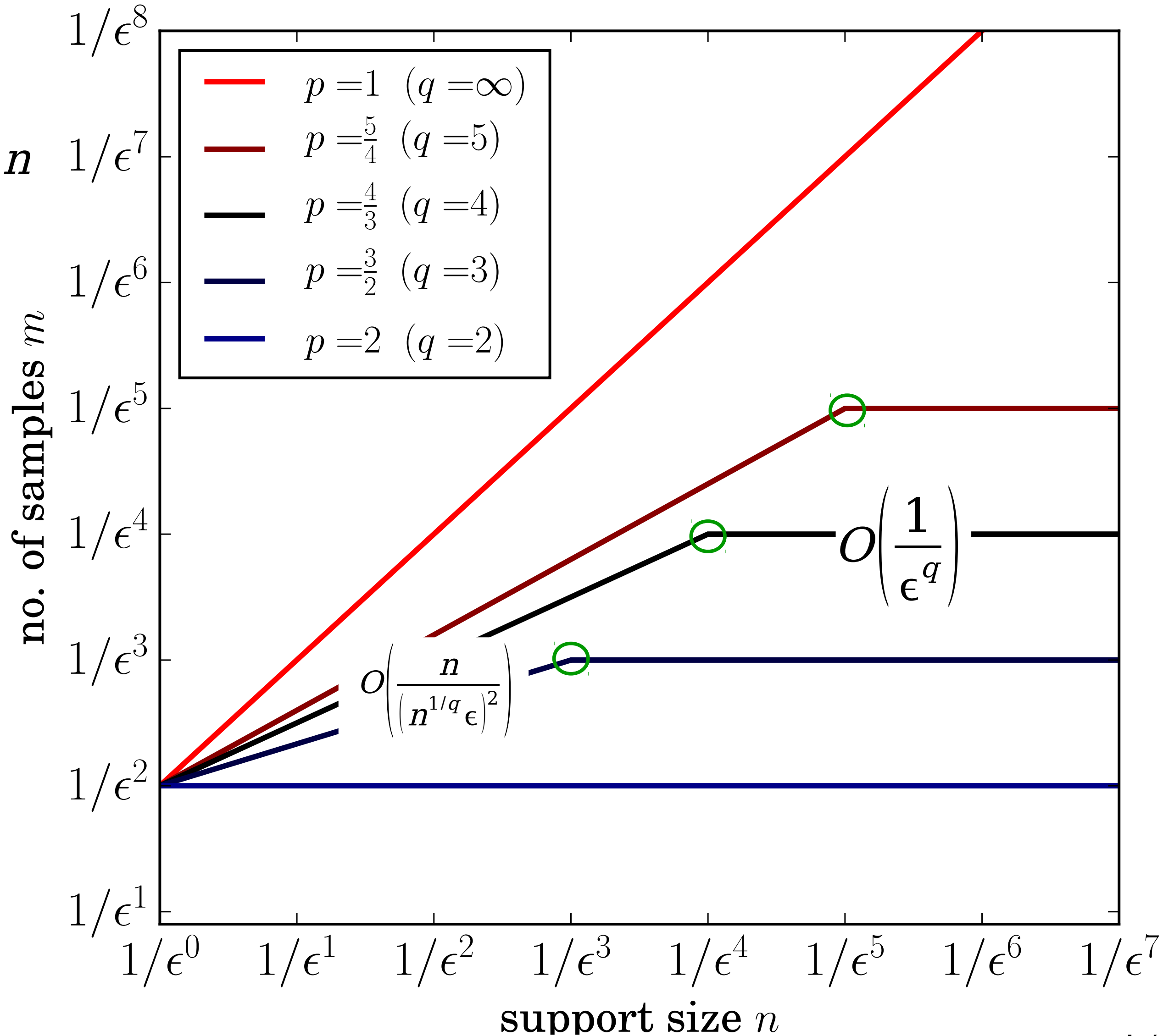
Learning

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

For $p > 1$:

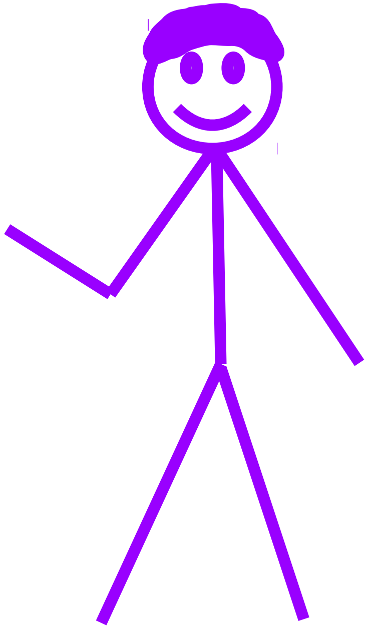
- Exists a sufficient # of samples independent of n
- Behavior differs in "small" and "large" n regimes

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



Outline

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

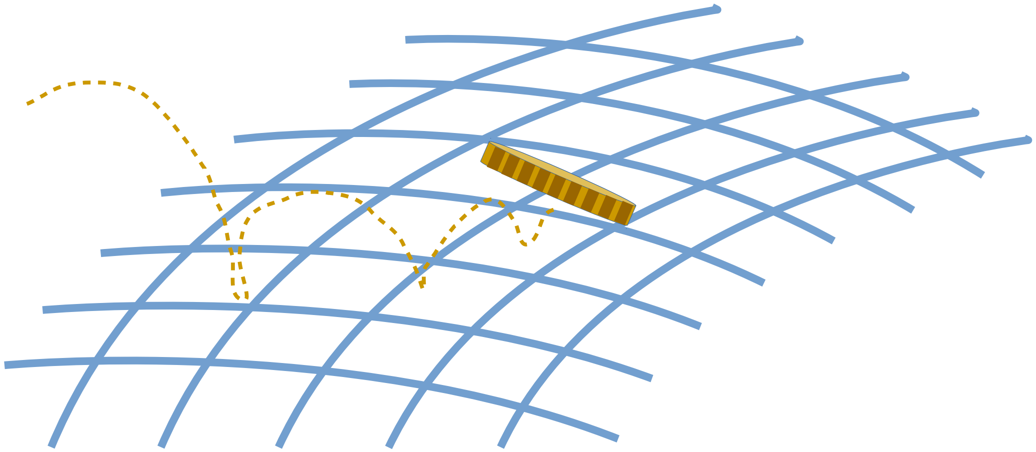


Classic Coin Question

Coin: either fair or one side with ϵ more probability.

Q: How many flips to tell?

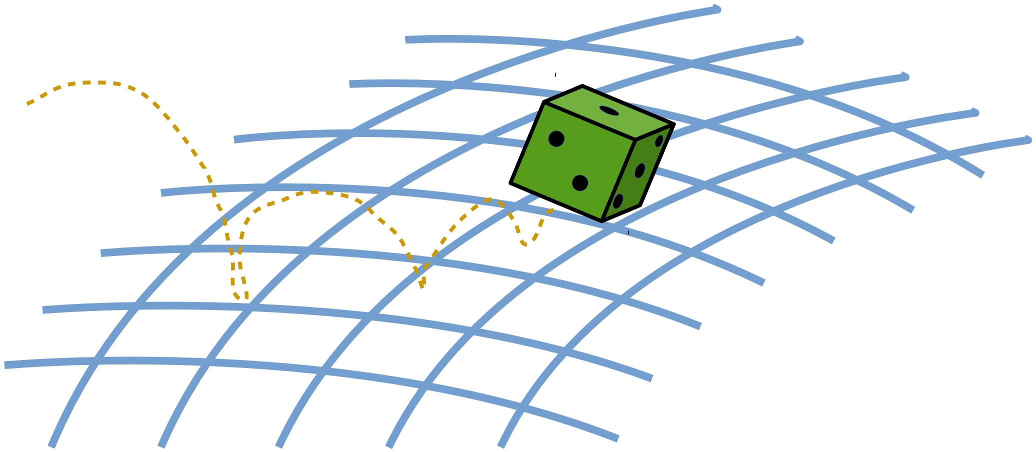
A: $O\left(\frac{1}{\epsilon^2}\right)$.



Classic Dice Question?

6-sided die: either fair or one side with ϵ more probability.

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

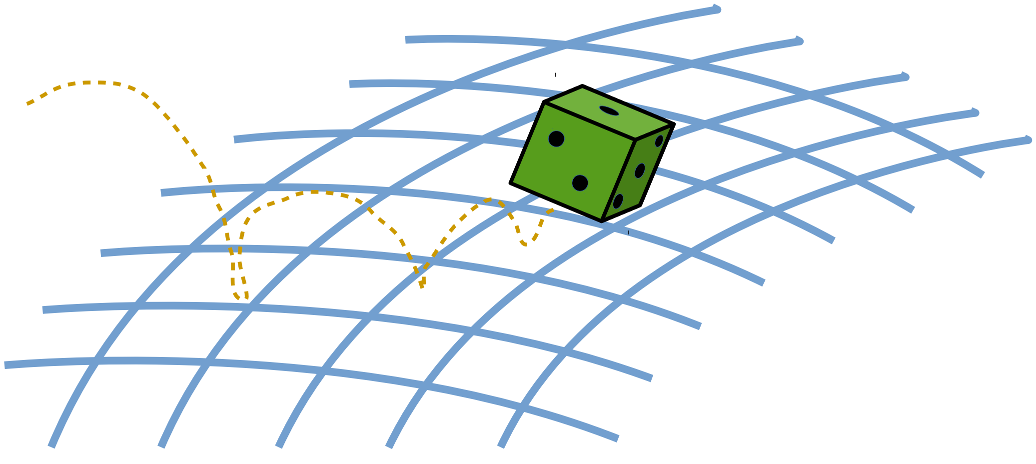


Classic Dice Question?

6-sided die: either fair or one side with ϵ more probability.

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

A: Fewer!

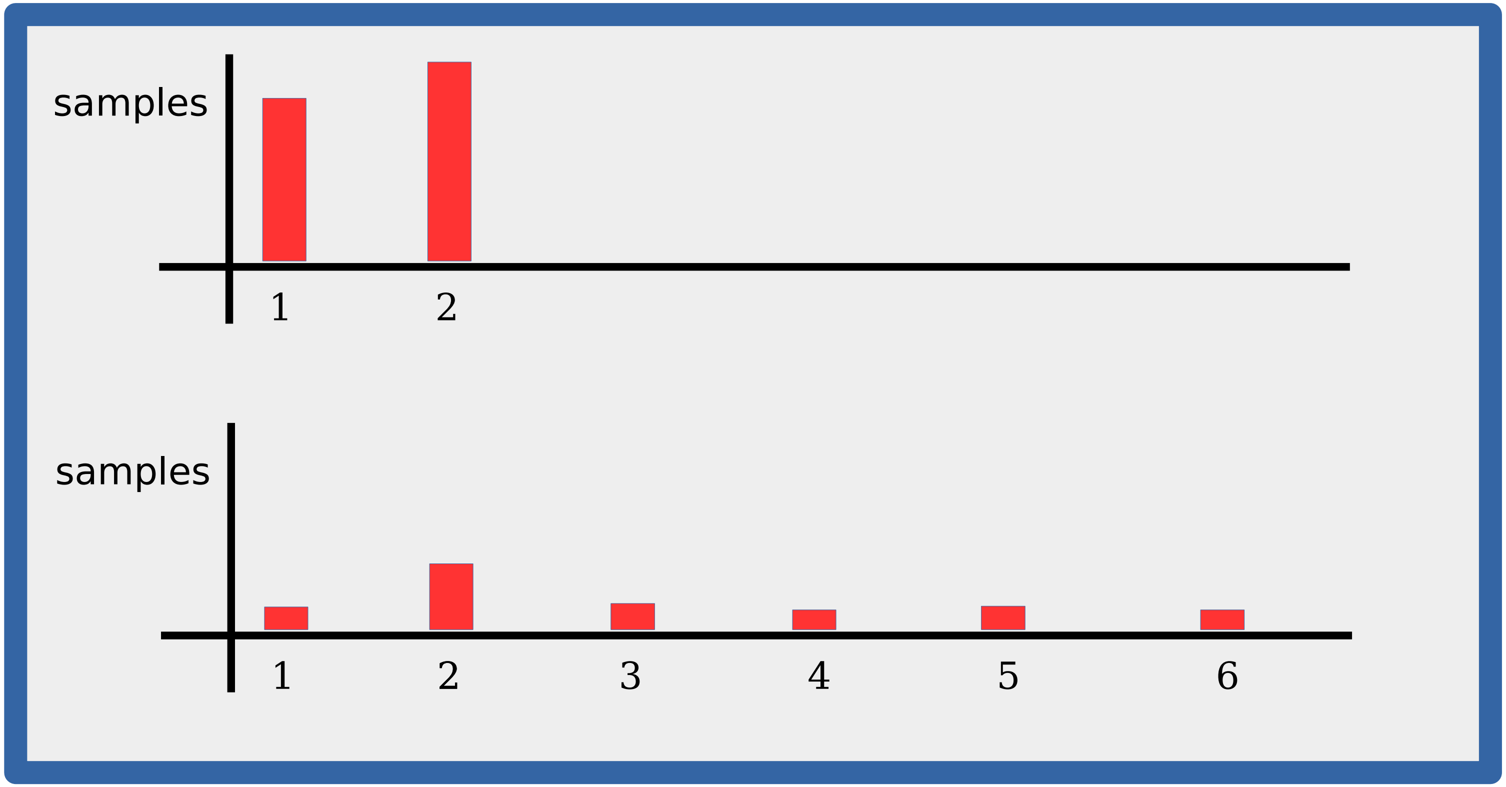


Classic Dice Question?

6-sided die: either fair or one side with ϵ more probability.

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

A: Fewer!



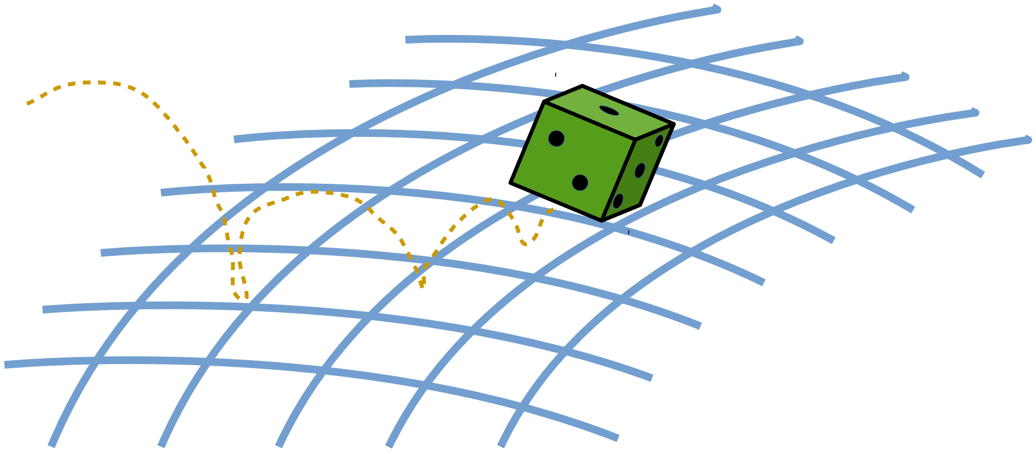
Classic Dice Question?

6-sided die: either fair or one side with ϵ more probability.

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

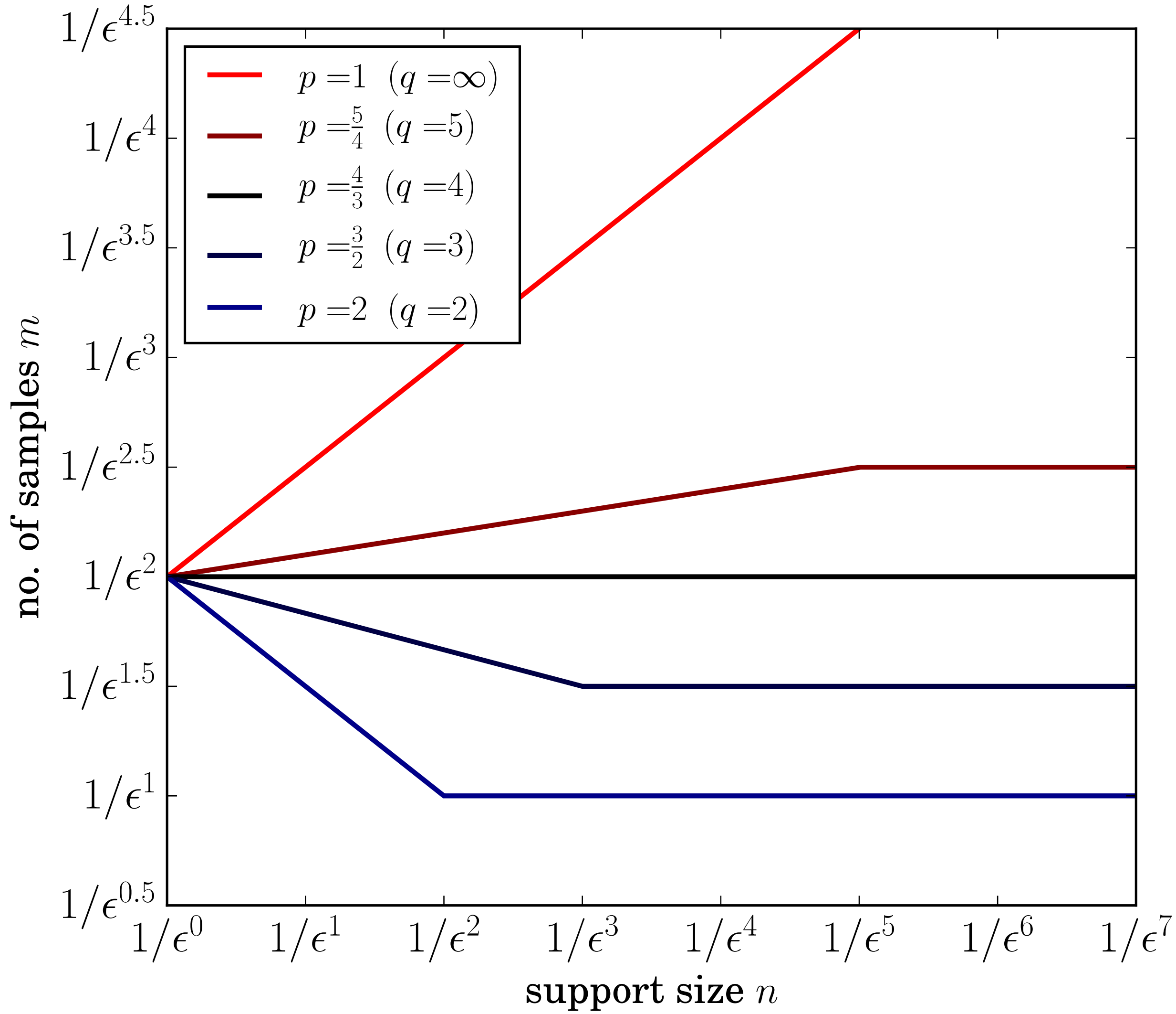
A: Fewer! (ℓ_∞)

For ℓ_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

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

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

Prior work counting collisions: Paninski (2008) (sort of); Goldreich and Don (2000); Batu, Fortnow, Rubinfeld, and Smith (2005)

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)

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):

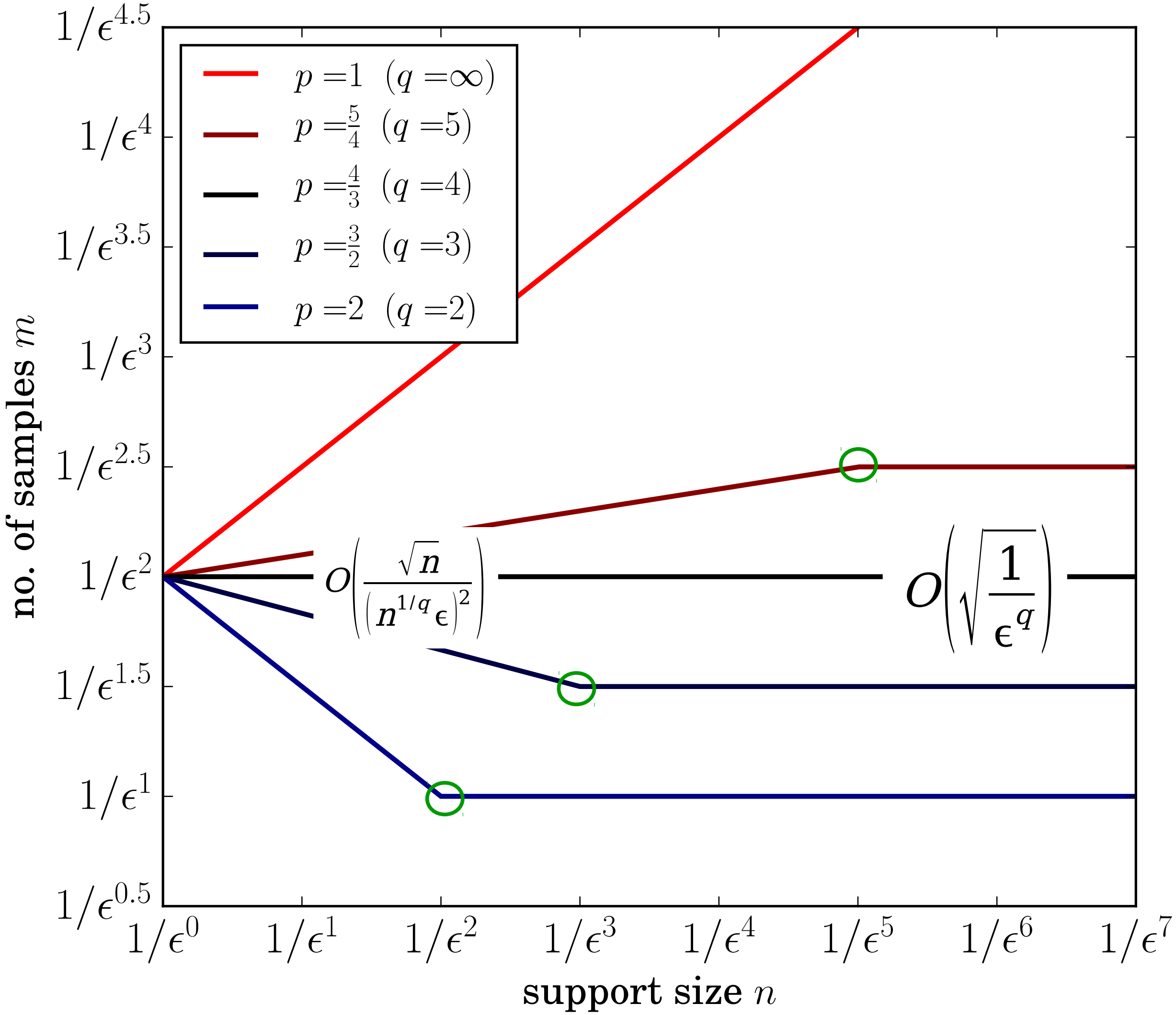
- For p = 1, $\frac{9}{\delta}\frac{\sqrt{n}}{\epsilon^2}$ samples are sufficient to test uniformity.

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

Testing, $1 \le p \le 2$

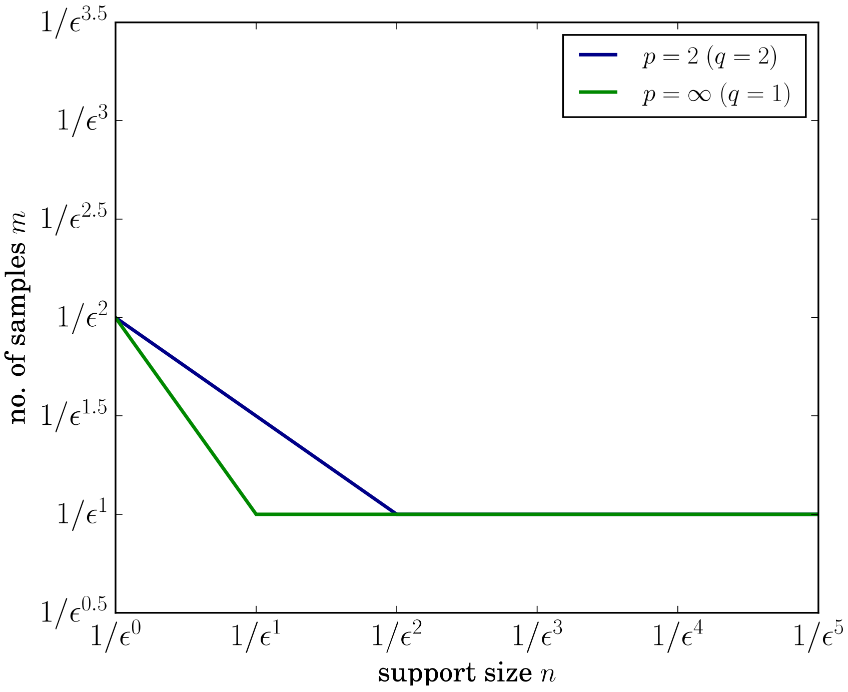
$$\|A-B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}}$$

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



ℓ_∞ Testing

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



ℓ_∞ Testing

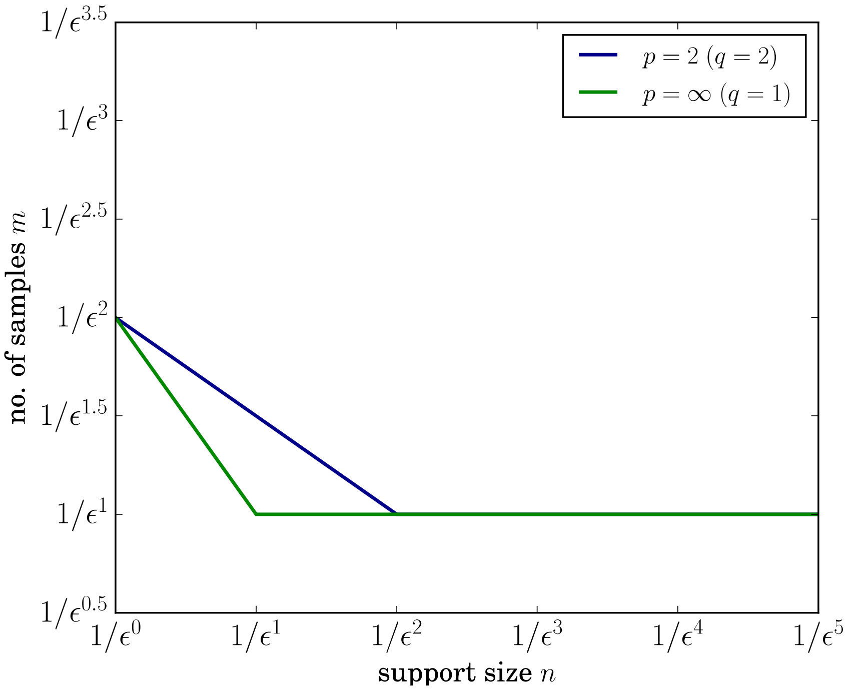
$$\|A-B\|_p = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}}$$

Theorem (for p = ∞):

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



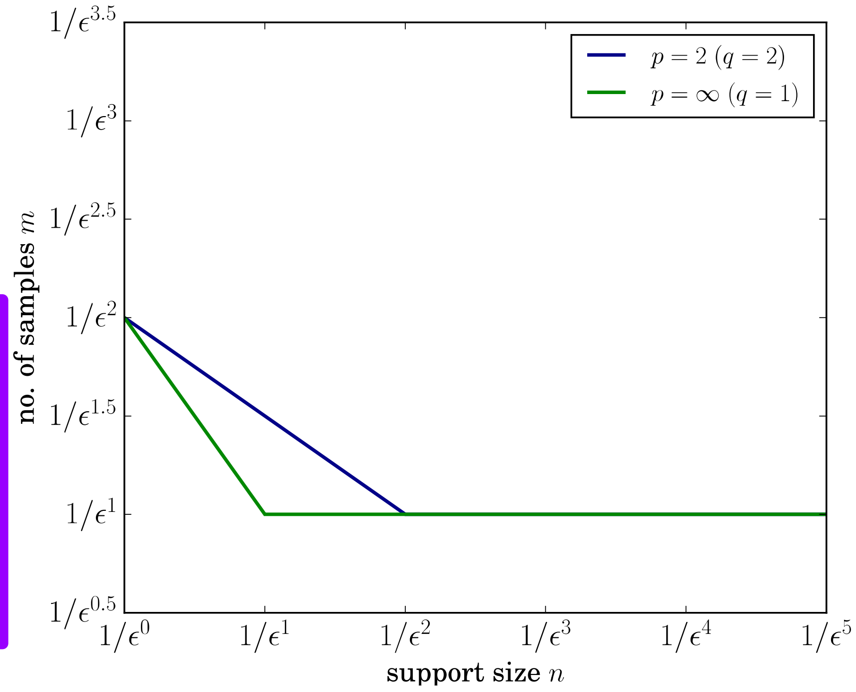
ℓ_∞ Testing

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

Theorem (for p = ∞):

- 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

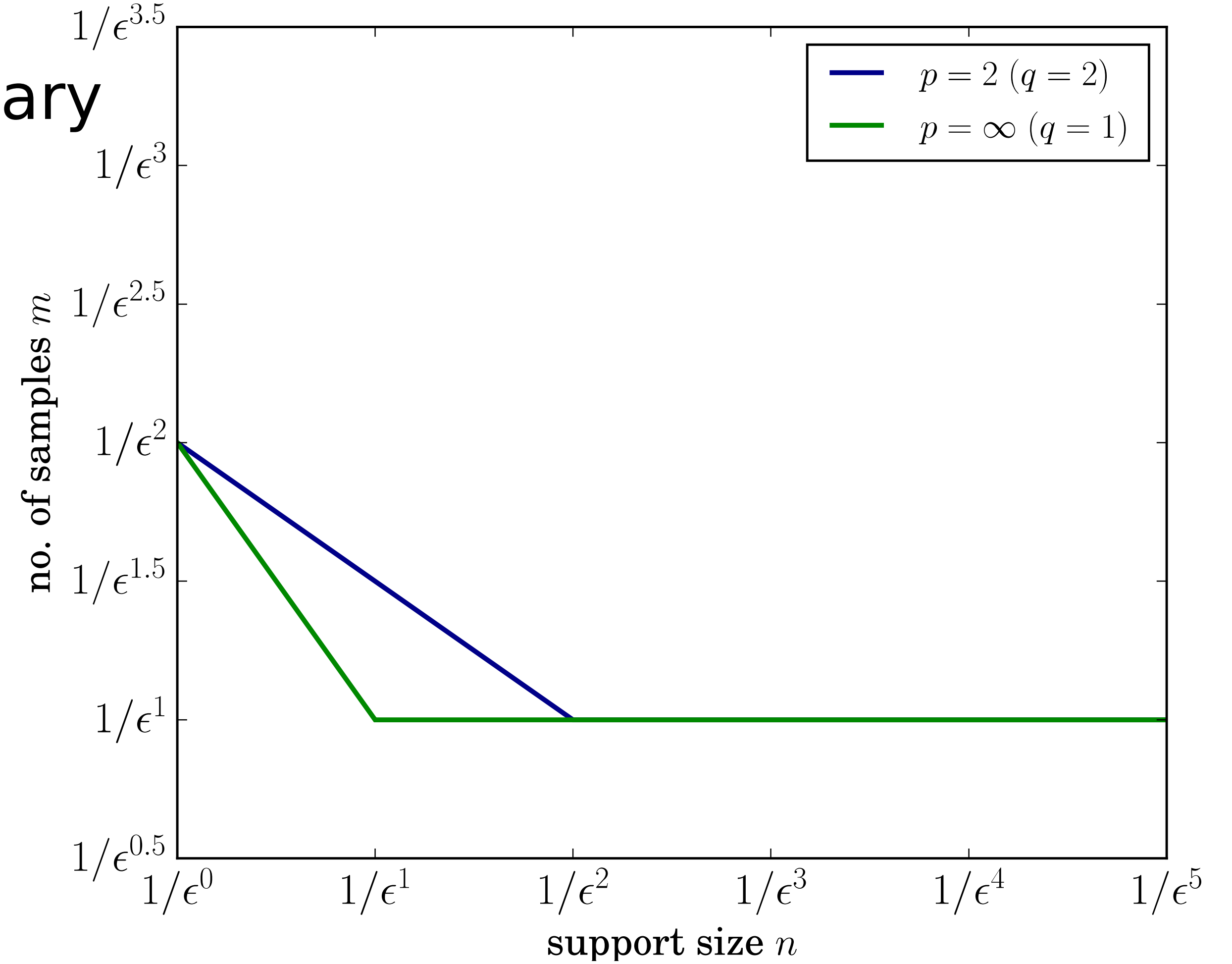
Gap for $2 < p < \infty$

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

- ℓ_2 alg $\rightarrow$ sufficient
  ℓ_∞ bound $\rightarrow$ necessary

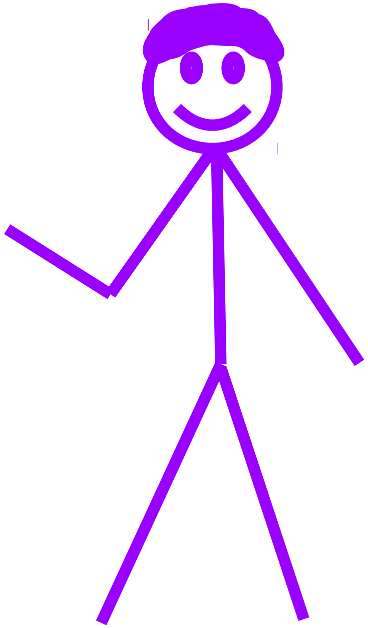
- Gap only in small-n case

- Seems to need different ideas



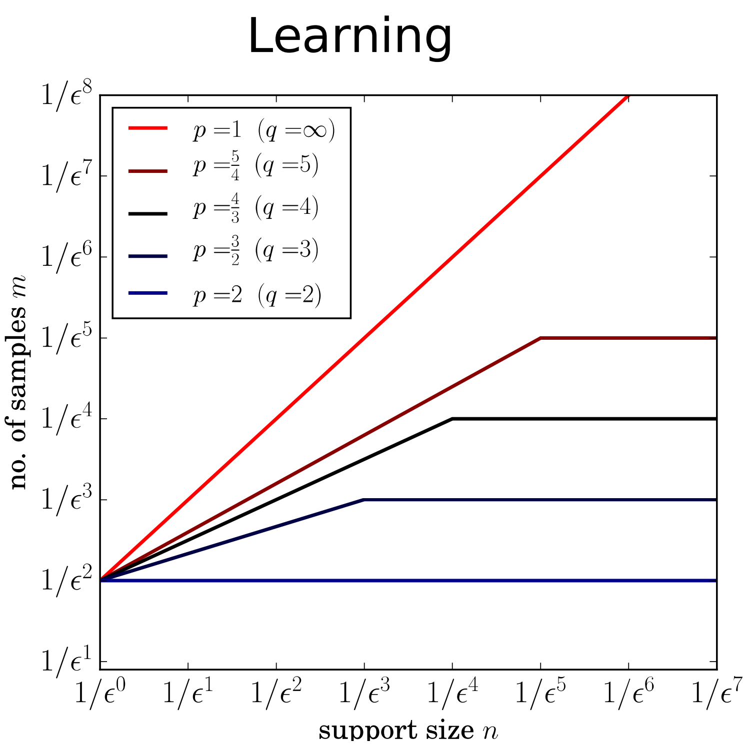
Outline

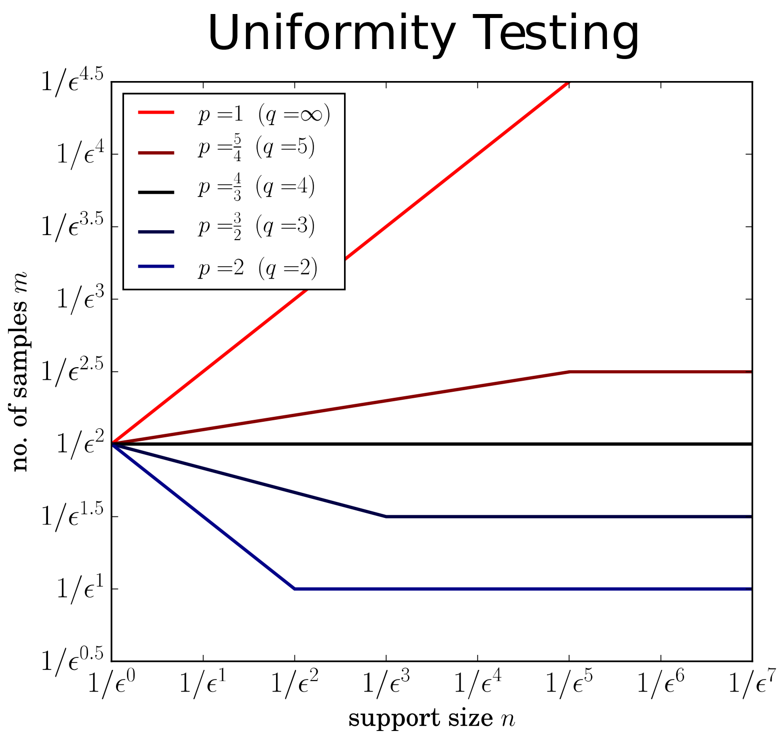
- Introductory stuff ✓
- Learning ✓
- Uniformity testing ✓
- Summary



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 ℓ_∞**: "almost-naive" alg is order-optimal



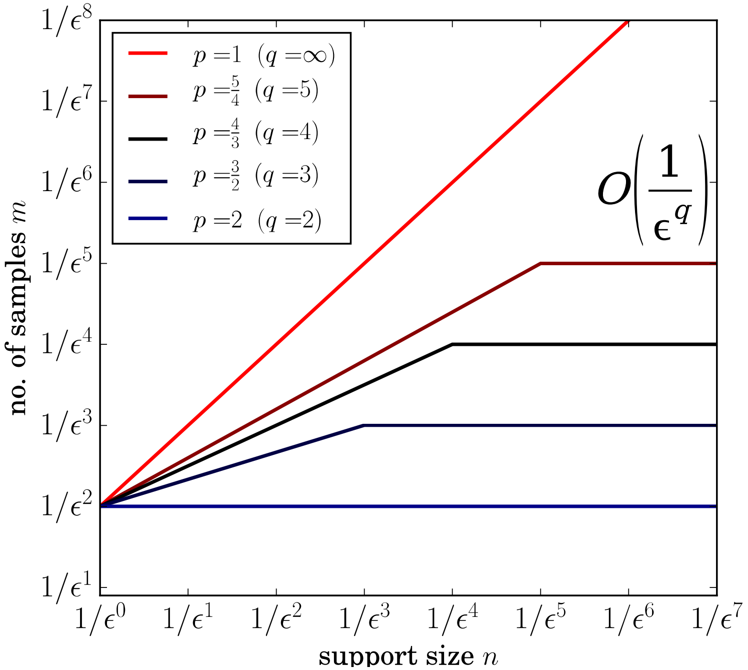


Ideas Summary

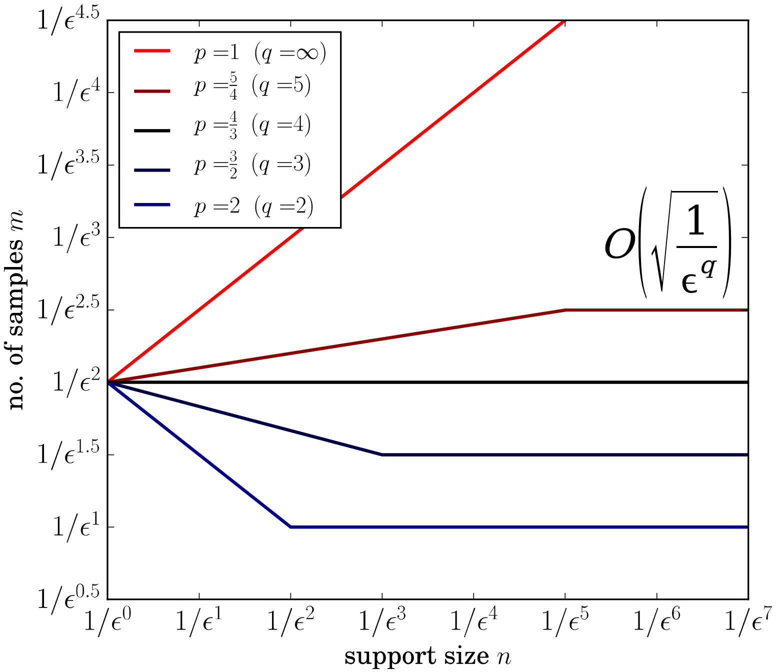
For $p > 1$:

- 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"

Learning



Uniformity Testing

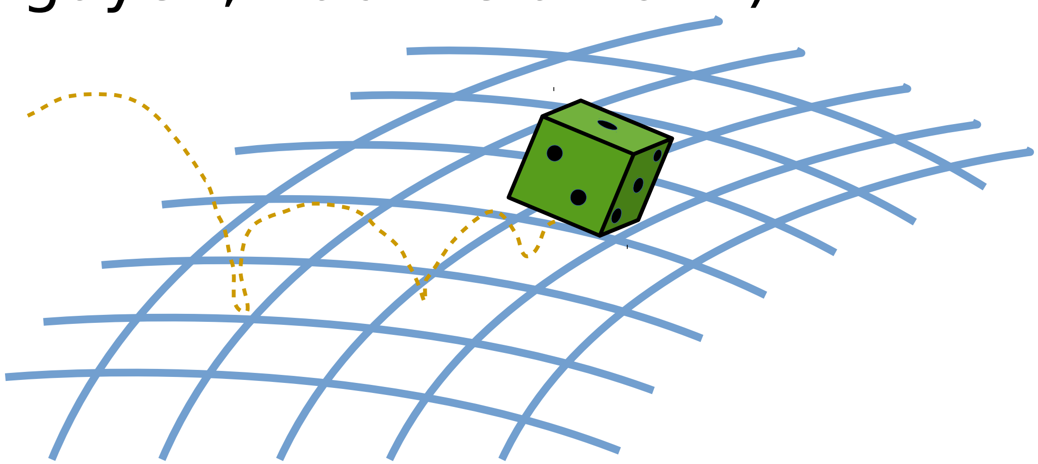


Future Work

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

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

- 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)

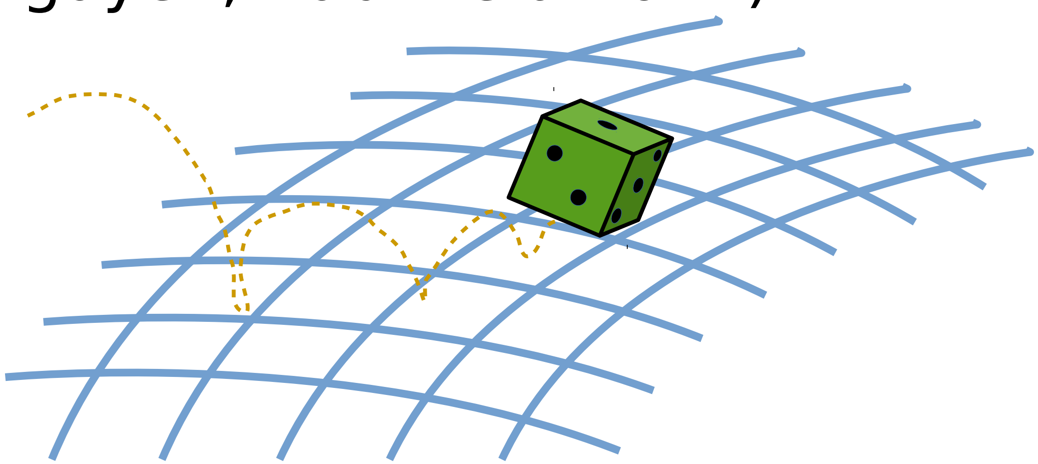


Future Work

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

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

- 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!