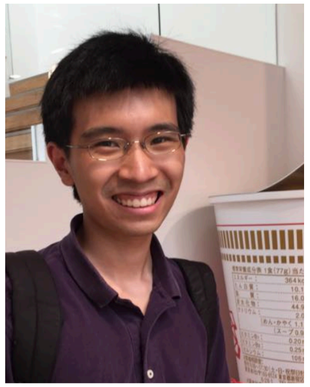
Interactive Proofs For Distribution Testing

Ari Biswas
University Of Warwick

Mark Bun

Boston University

Clément Canonne

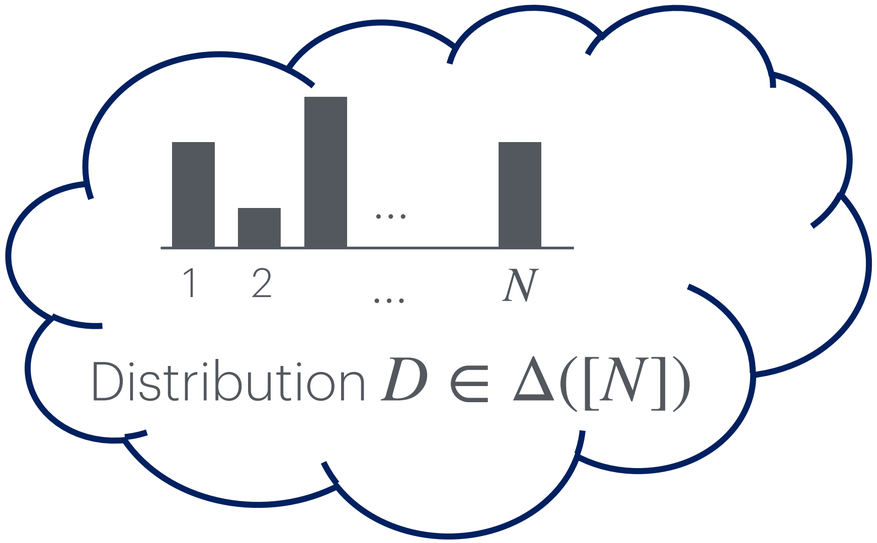

University Of Sydney

Satchit Sivakumar

Boston University

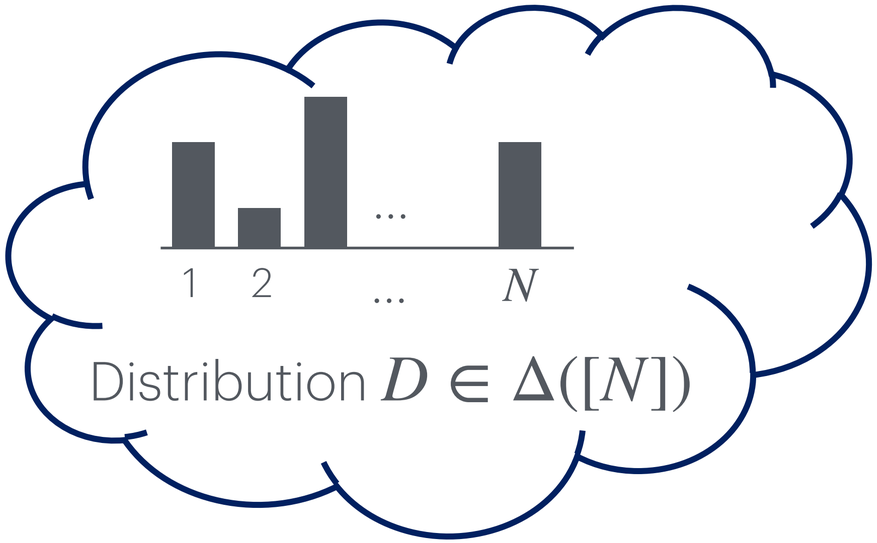
Speaker

Distribution Testing - Setup



Testing Algorithm

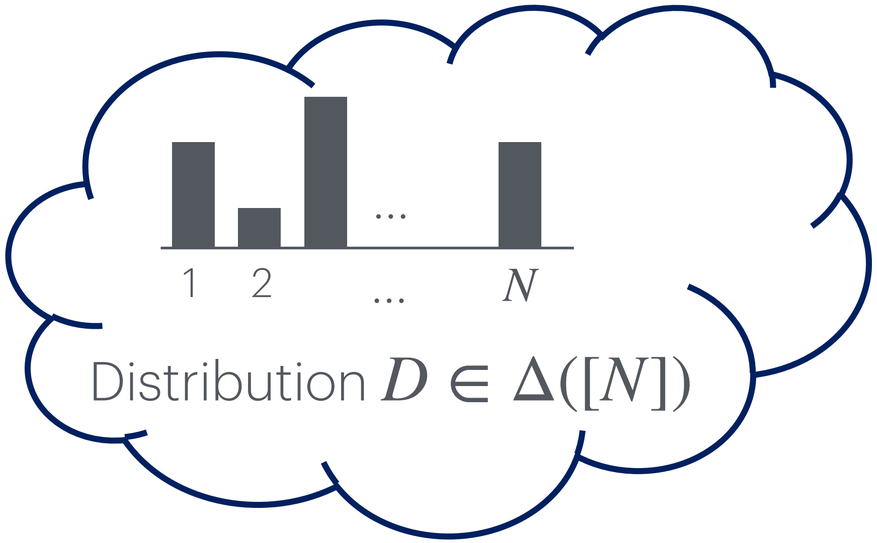
Distribution Testing - Setup



m i.i.d samples

Testing Algorithm

Distribution Testing - Setup



m i.i.d samples

Testing Algorithm

From just the samples alone, the tester wants to decide if D belongs to some property

Examples of Properties:

- **Is D (approximately) uniform?**
- **Is the support size of D greater than 100 ?**
- **Is the Shannon entropy of D less than $\log^2 N$?**
- **Is the PMF of D monotone?**

Distribution Testing - More Formally

[Batu-Fortnow-Rubinfeld-Smith-White, 2000]

Testing Algorithm

- A **property** is a set $\Pi \subseteq \Delta([N])$ of distributions.
- The Decision Problem:

Given sample access to an unknown distribution D and a description of a property Π, a testing algorithm A must do the following:

Completeness: If $D \in \Pi$, $\Pr[A^{(D)}\text{accpets}] \geq 2/3$

Soundness: If D is ε-far from every distribution in Π, $\Pr[A^{(D)}\text{accpets}] \leq 1/3$

Complexity: The number of samples A draws from D

Distribution Testing - More Formally

orithm

- A **property** is a set $\Pi \subseteq \Delta([N])$ of distributi
- The Decision Problem:

G
te

Cor ccpets] $\geq 2/3$

Soun D is ε-far from every distribution in Π, $\Pr[A^{(D)}\text{accpets}] \leq 1/3$

The Game: How many samples are necessary/sufficient to decide property Π

Complexity:

The number of samples A draws from D

Some Known Results

Baseline :Can test ***any*** property with $m = O(N/\varepsilon^2)$ samples (just learn the distribution)

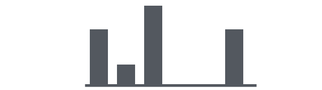
- So for the problem to be practical/interesting -- want to be sub-linear in N.
- $\Pi =$ Uniform Distribution, $m = \theta(\sqrt{N})$ [Goldreich-Ron, 2000]
- Π = Distributions with Shannon entropy greater than k, $m = \theta(N/\log N)$ samples [Raskhodnikova-Ron-Sppilka-Smith, 2007, Valiant 2008, Valiant-Valiant 2011]

  Any label-invariant property actually

- Π = Distributions with monotone mass functions over $[N]$ [Batu-Kumar-Rubinfeld04, Bhattacharyya-Fischer-Rubinfeld-Valiant10, Acharya-Daskalakis-Kamath15, Canonne-Diakonikolas-Gouleakis-Rubinfeld16, Aliakbarpour-Gouleakis-Peebles-Rubinfeld-Yodpinyanee19]
  - $\theta(\sqrt{N})$: Total Ordering
  - $\Omega(N/\log N)$: Partial Ordering

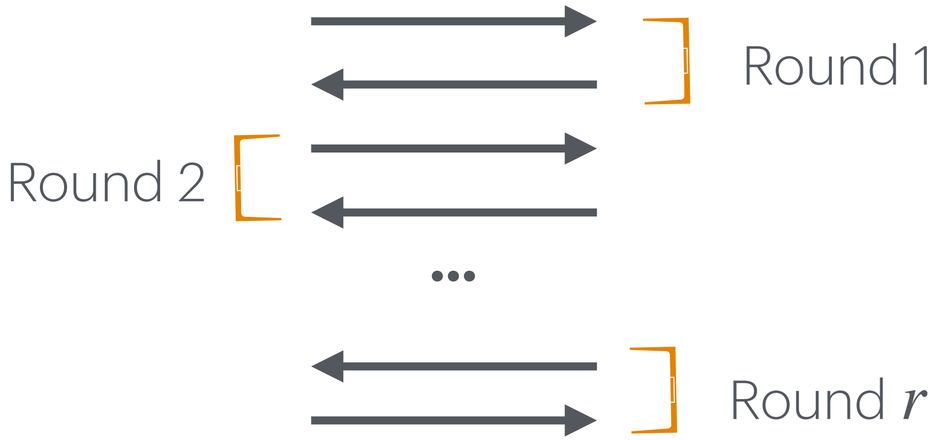
Proofs For Distributions

[Chiesa-Gur 2018]

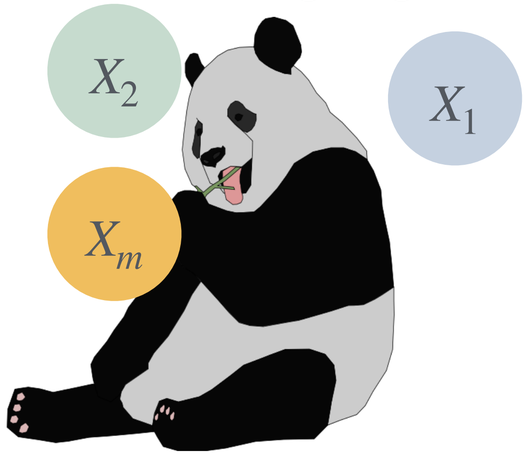


Untrusted Prover with complete knowledge of the mass of the distribution

A proof π is a prescribed next message protocol



Testing Algorithm

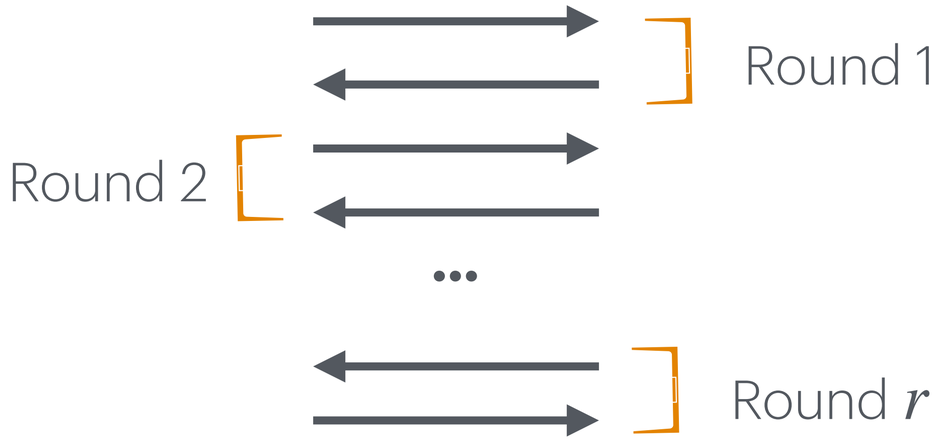


Proofs For Distributions

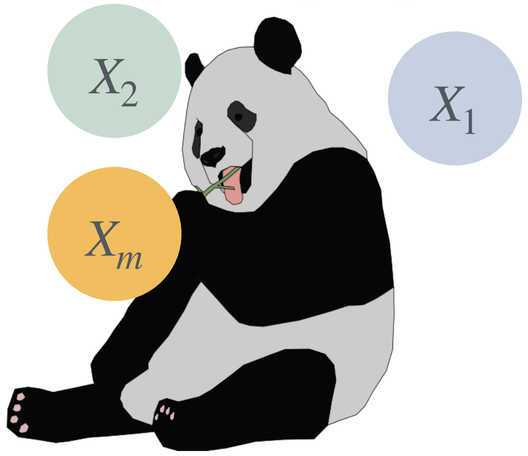
[Chiesa-Gur 2018]

Prover is doubly efficient when it needs to learn the distribution by sampling

A proof π is a prescribed next message protocol



Testing Algorithm



But allowed to take $O(\text{poly}(N))$ samples

Proofs For Distributions

- A **property** is a set $\Pi \subseteq \Delta([N])$ of distributions.

- **<u>The Decision Problem:</u>** Given sample access to an unknown distribution D, a description of a property Π and ability to exchange messages with a prover , we want to come up with a proof π such that

<u>Completeness</u>: If $D \in \Pi$, and the prover follows π, then $\Pr[A^{(D)}\text{accpets}] \geq 2/3$

<u>Soundess</u>: If D is ε-far, and then for **<u>all</u>** provers, then $\Pr[A^{(D)}\text{accpets}] \leq 2/3$

Proofs For Distributions

- **The Decision Problem:** Given sample access to an unknown distribution D, a description of a property Π and ability to exchange messages with a prover , we want to come up with a proof π such that

Completeness: If $D \in \Pi$, and the prover follows π, then $\Pr[A^{(D)}\text{accpets}] \geq 2/3$

Soundess: If D is ε-far, and then for **all** provers, then $\Pr[A^{(D)}\text{accpets}] \leq 2/3$

Sample Complexity**:** Number of samples drawn by testing algorithm

Communication Complexity**:** Number of bits communicated by the prover and tester

Round Complexity**:** Number of rounds of interaction

Proofs For Distributions

- A **property** is a set $\Pi \subseteq \Delta([N])$ of distributions.

- **<u>The Decision Problem:</u>** Given sample access to an unknown distribution D, and ability to exchange messages with a prover , we want to come up with a proof π such that

Other Nice Properties:

1. Tolerance: Accept whenever D is close to Π
2. Prover is doubly efficient
3. 1 Round communication or no rounds
4. Public coin -- everything the verifier samples is revealed to the prover
   i.e. verifier has no secrets

Round Complexity**<u>:</u>** Number of rounds of interaction

Known Results

[Chiesa-Gur, 2018]

Baseline :Can test *any* property with $m = O(\sqrt{N})$ samples and $\widetilde{O}(N)$ communication

STEP 1
Prover computes
$\hat{D}$ = discretisation of D to error ε/N, described in $\widetilde{O}(N)$ bits, and communicates this to the tester/verifier.

(Tolerant) Identity Testing $\Longrightarrow$ Verification

STEP 2: Using $O(\sqrt{N})$ samples, determine whether

$$\mathrm{TV}(\hat{D}, D) \leq \varepsilon/N$$
or
$$\mathrm{TV}(\hat{D}, D) \geq \varepsilon$$

(Tolerant) Identity Testing

$\mathrm{TV}(\hat{D}, D) \leq \varepsilon_1$
or
$\mathrm{TV}(\hat{D}, D) \geq \varepsilon_2$

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right),$$

[Canonne-Jain-Kamath-Li, 2021]

Known Results

[Chiesa-Gur, 2018]

Baseline :Can test *any* property with $m = O(\sqrt{N})$ samples and $\widetilde{O}(N)$ communication

STEP 1
Prover computes
$\hat{D}$ = discretisation of D to error ε/N, described in $\widetilde{O}(N)$ bits, and communicates this to the tester/verifier.

(Tolerant) Identity Testing $\Longrightarrow$ Verification

STEP 2: Using $O(\sqrt{N})$ samples, determine whether

$$\mathrm{TV}(\hat{D}, D) \leq \varepsilon/N$$
or
$$\mathrm{TV}(\hat{D}, D) \geq \varepsilon$$

(Tolerant) Identity Testing
$\mathrm{TV}(\hat{D}, D) \leq \varepsilon_1$
or
$\mathrm{TV}(\hat{D}, D) \geq \varepsilon_2$

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right),$$

[Canonne-Jain-Kamath-Li, 2021]

Sub-linear Communication is the interesting regime

Known Results: Without Provers

- $\Pi =$ Uniform Distribution, $m = \theta(\sqrt{N})$ [Goldreich-Ron, 2000]

- Π = Distributions with Shannon entropy greater than k, we need $m = \theta(N/\log N)$ samples
  Any label-invariant property actually
  [Raskhodnikova-Ron-Sppilka-Smith, 2007,
  Valiant 2008
  Valiant-Valiant 2011]

- Π = Distributions with monotone mass functions over $[N]$
  - $\Omega(N/\log N)$: Partial Ordering

[Batu-Kumar-Rubinfeld04, Bhattacharyya-Fischer-Rubinfeld-Valiant10, Acharya-Daskalakis-Kamath15, Canonne-Diakonikolas-Gouleakis-Rubinfeld16, Aliakbarpour-Gouleakis-Peebles-Rubinfeld-Yodpinyanee19

Known Results: With A Prover

[Goldreich-Ron, 2000], [Herman-Rothblum 22,]

- $\Pi =$ Uniform Distribution, $m = \theta(\sqrt{N})$ with no communication from prover. $\Omega(\sqrt{N})$ samples regardless of communication.

[Herman-Rothblum 22, 23]

- Π = Distributions with Shannon entropy greater than k, ~~$m = \theta(N/\log N)$~~ $m = \tilde{O}(\sqrt{N})$ samples and communication *Any label-invariant property actually*

- Π = Distributions with monotone mass functions over $[N]$
  - $O(N^{1-\alpha})$, for some $\alpha > 0$ [Herman-Rothblum 24] *Essentially all properties we care about*

Known Results: With A Prover

[Goldreich-Ron, 2000],

[Herman-Rothblum 22,]

- $\Pi =$ Uniform Distribution, $m = \theta(\sqrt{N})$ with no communication from prover. $\Omega(\sqrt{N})$ samples regardless of communication

[Herman-Rothblum 22, 23]

$\Omega(\sqrt{N})$ $m = \tilde{O}(\sqrt{N})$

Rough Idea is the following -- If the prover were to **honestly** communicate the entire mass function of the distribution to the verifier, the verifier is just as powerful as the prover. The problem of verifying now reduces to identity testing, for which we have lower bounds of $\sqrt{N}$ samples. If the verifier could beat this, then it would violate these known lower bounds.

care about

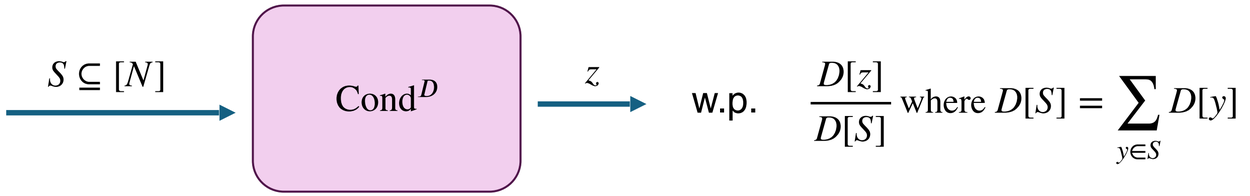
Towards Practical Verification

[Chakraborty-Fischer-Goldhirsh-Matsliah13, Canonne-Ron-Servedio14]

Question: How might we augment the verifier's power so that it can get away with $\ll \sqrt{N}$ samples?

This work: Give the verifier *conditional* queries

Verifying Algorithm



Towards Practical Verification

[Chakraborty-Fischer-Goldhirsh-Matsliah13, Canonne-Ron-Servedio14]

Question: How might we augment the verifier's power so that it can get away with $\ll \sqrt{N}$ samples?

This work: Give the verifier *conditional* queries

[Chakraborty-Fischer-Goldhirsh-Matsliah13]

cf. In the full $\boldsymbol{COND}$ query model, every label-invariant property can be tested with $\mathbf{polylog}\, N$ queries

Verifying Algorithm

$S \subseteq [N]$ → Cond^D → z w.p. $\dfrac{D[z]}{D[S]}$ where $D[S] = \sum_{y \in S} D[y]$

Towards Practical Verification

Pairwise Conditional Queries

Special (more realistic?) case of conditional queries where $\left|S\right| = 2$

<u>To draw random samples:</u>
Survey everyone and ask for their occupation

<u>To draw pairwise conditional samples:</u>
Target survey to, e.g., “nurses or firefighters”

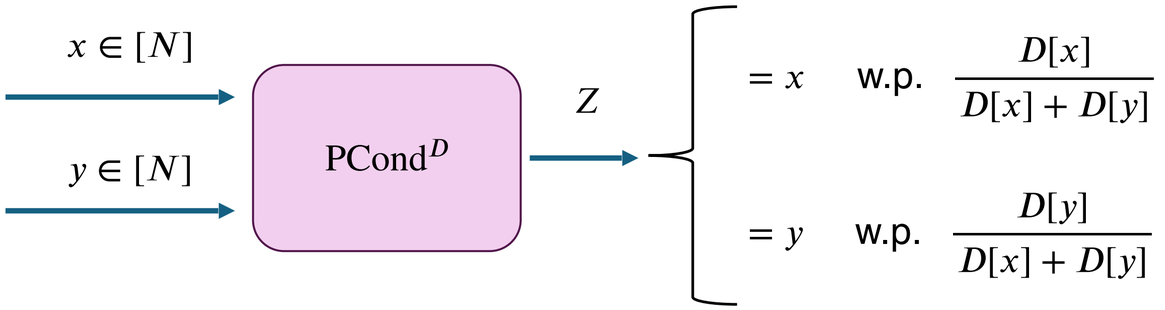
Towards Practical But Reasonable Verification

[Chakraborty-Fischer-Goldhirsh-Matsliah13, Canonne-Ron-Servedio14]

Question: How might we (slightly) augment the verifier's power so that it can get away with $\ll \sqrt{N}$ samples?

This work: Give the verifier ***pairwise*** *conditional* queries

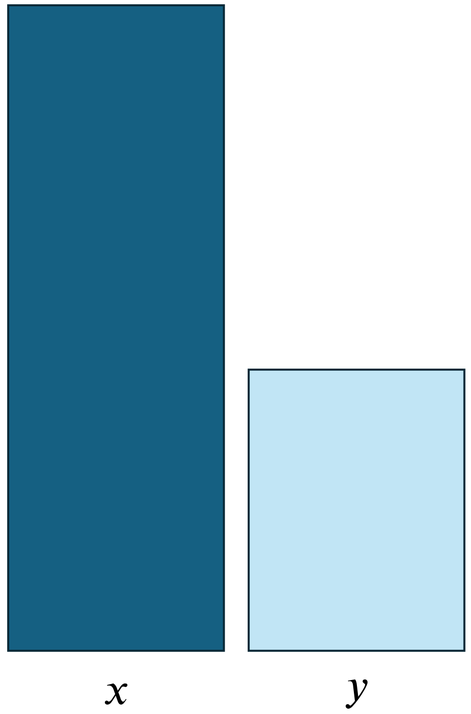
Verifying Algorithm



Pairwise Conditional Queries

Appear Harmless But Quite Powerful

Warmup: Can compare the probability mass on two arbitrary points



Using $O\left(\frac{K}{\eta^2}\right)$ $PCOND^D$ queries to $(x,\ y)$, can output either

1. An estimate of $\frac{D[x]}{D[y]}$ to within multiplicative $(1+\eta)$, or

2. Signal that either $\frac{D[x]}{D[y]} > K$ or $\frac{D[x]}{D[y]} < \frac{1}{K}$

[Canonne-Ron-Servedio14]

Pairwise Conditional Queries

Appear Harmless But Quite Powerful

Theorem: Can ε-test uniformity using $O(1/\varepsilon^4)$ PCOND queries + samples [Canonne-Ron-Servedio14]

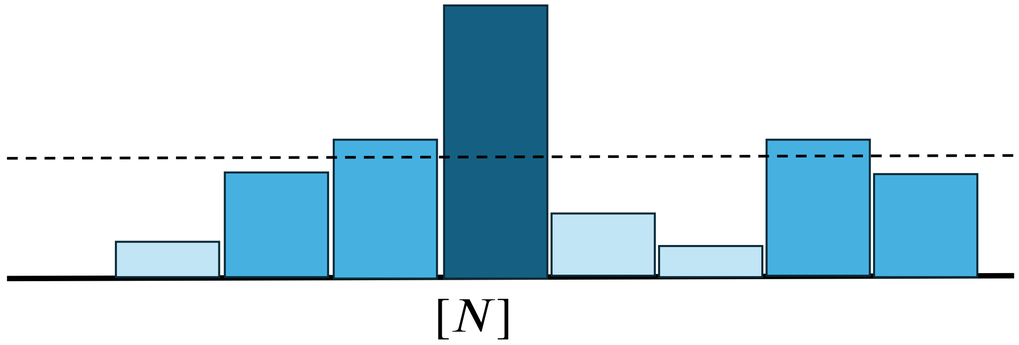
cf. testing, and even verifying, uniformity with just samples requires $\Omega(\sqrt{N})$ samples

1. Draw $O\left(\frac{1}{\varepsilon}\right)$ samples H from D
   Whp, at least one is "heavy"

2. Draw $O\left(\frac{1}{\varepsilon}\right)$ samples L from Unif_N
   Whp, at least one is "light"

3. Compare each pair $x \in H$, $y \in L$ using $O\left(\frac{1}{\varepsilon^2}\right)$ PCOND queries

Idea: If D is far from uniform, it must have lots of mass on "heavy" elements and lots of "light" elements



Pairwise Conditional Queries

Appear Harmless But Quite Powerful

Theorem: Can ε-test uniformity using $O(1/\varepsilon^2)$ PCOND queries + samples [Canonne-Ron-Servedio14]

Theorem: Can ε-test identity, i.e., equivalence to any fixed distribution D^*, using $O(\sqrt{\log N}/\varepsilon^2)$ PCOND queries + samples [Canonne-Ron-Servedio14, Narayanan 21]

cf. testing, and even verifying, uniformity with just samples requires $\Omega(\sqrt{N})$ samples

Theorem: Can estimate the mass in "neighbourhood" of a point whenever it's sufficiently large

i.e., if $D\left[U(y)\right] \geq \beta$, can estimate to within multiplicative $(1+\eta)$

using $O\left(\frac{1}{\varepsilon^2\eta^4\beta^2}\right)$ queries + samples

where $U(y) = \{z \in [N] : D[z] \in (1 \pm \varepsilon)D[y]\}$

Pairwise Conditional Queries

Appear Harmless But Quite Powerful

Theorem: Can ε-test uniformity using $O(1/\varepsilon^2)$ PCOND queries + samples [Canonne-Ron-Servedio14]

Theorem: Can ε-test identity, i.e., equivalence to any fixed distribution D^*, using $O(\sqrt{\log N}/\varepsilon^2)$ PCOND queries + samples [Canonne-Ron-Servedio14, Narayanan 21]

cf. testing, and even verifying, uniformity with just samples requires $\Omega(\sqrt{N})$ samples

Theorem: Can estimate the mass in "neighbourhood" of a point whenever it's sufficiently large

i.e., if $D\left[U(y)\right] \geq \beta$, can estimate to within multiplicative $(1+\eta)$

using $O\left(\frac{1}{\varepsilon^2\eta^4\beta^2}\right)$ queries + samples

where $U(y) = \{z \in [N] : D[z] \in (1 \pm \varepsilon)D[y]\}$

This work: Can also sample from $U(y)$

Pairwise Conditional Queries

Appear Harmless But Quite Powerful; But Not That Powerful

This work: There is a label-invariant property Π which requires at least $\Omega(N^{1/8})$ PCOND queries + samples to test

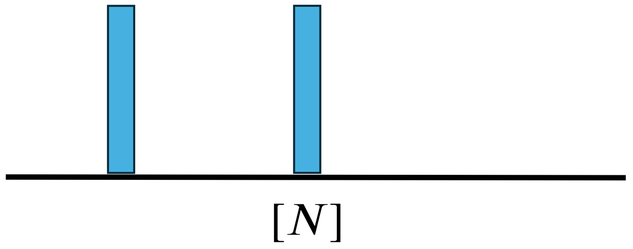
$$\Pi = \{D \text{ is uniform over exactly } N^{1/4} \text{ domain elements}\}$$

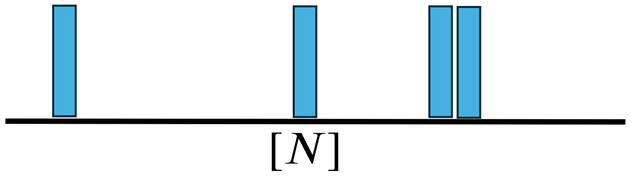
[Chakraborty-Fischer-Goldhirsh-Matsliah13]

cf. In the full $COND$ query model, every label-invariant property can be tested with $\text{polylog}N$ queries

Proof idea: Distributions in Π are hard to distinguish from those in $\Pi' = \{D \text{ is uniform over exactly } N^{1/2} \text{ domain elements}\}$ since

- $o(N^{1/8})$ samples is too few to get collisions
- PCOND queries are only helpful when they involve points in the support of D, but not in the sample...but these are hard to find!



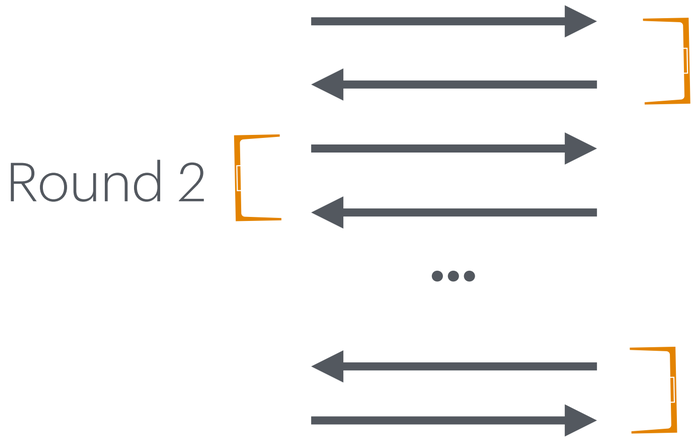


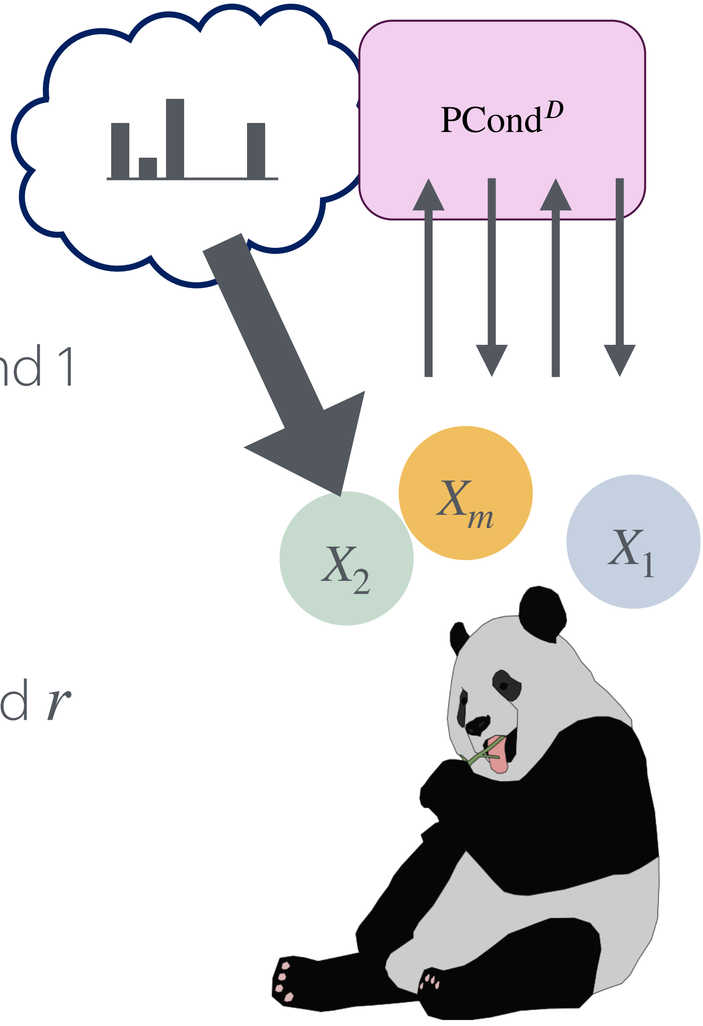
Verification With P-COND queries

THIS WORK

Untrusted Prover with complete knowledge of the mass of the distribution

A proof π is a prescribed next message protocol

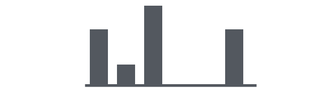




Testing Algorithm

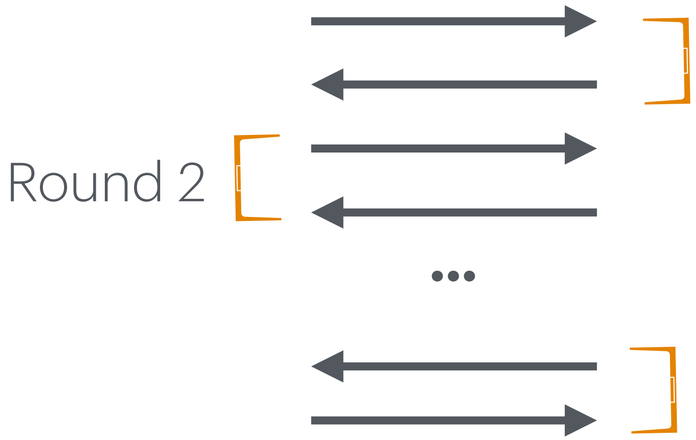
Verification With P-COND queries

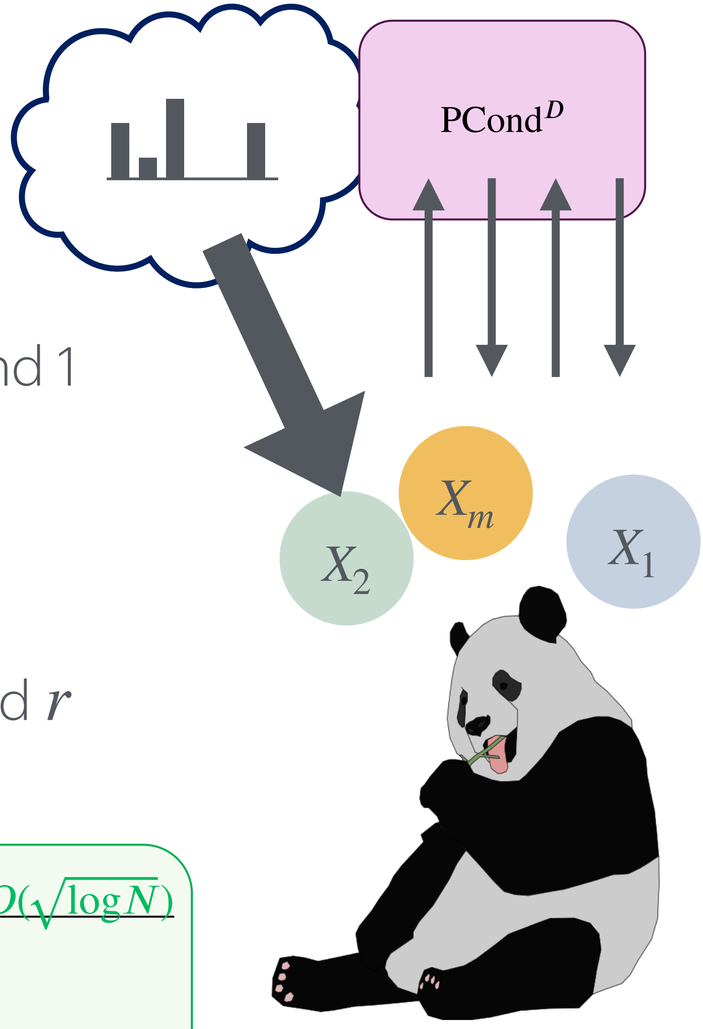
THIS WORK



Untrusted Prover with complete knowledge of the mass of the distribution

A proof π is a prescribed next message protocol





Testing Algorithm

Baseline: Every property Π can be verified with $O(\sqrt{\log N})$ queries + samples and $\tilde{O}(N)$ communication

[Chiesa-Gur18 + Narayanan21]

Main Result

Every label-invariant property Π of distributions on $[N]$ can be verified using

- poly$(\log N)$ samples,
- poly$(\log N)$ **PCOND** queries, and
- $\tilde{O}\left(\sqrt{N}\right)$ communication (over poly$(\log N)$ rounds of interactions)

Proof Overview

Every label-invariant property can be verified using

- $\mathrm{polylog}\, N$ samples,
- $\mathrm{polylog}\, N$ PCOND queries, and
- $\tilde{O}\left(\sqrt{N}\right)$ communication

1. Verifying a label-invariant property

2. Verifying a bucket histogram

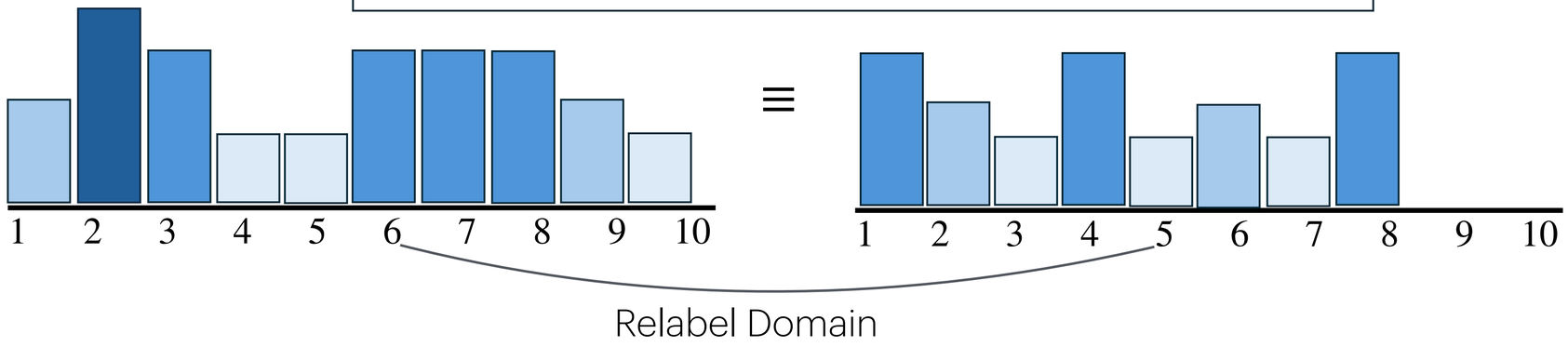
3. Verifying the probability mass of one point

4. Verifying the support size of a nearly-flat distribution

Bucket Histograms

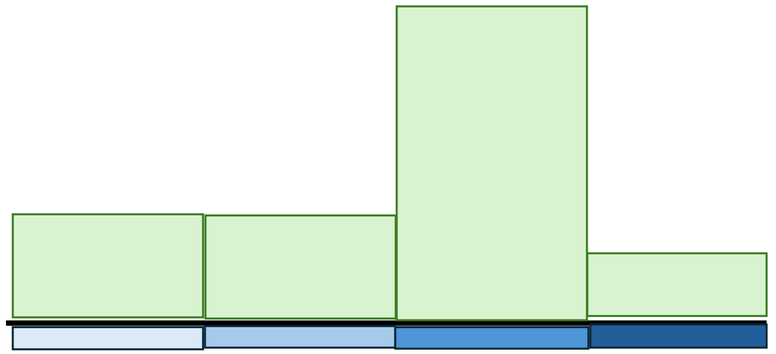
A label-invariant property depends only on a distribution's weights, not on the underlying elements

probability mass function



Both distributions have the same histogram

total mass on elements with given weight



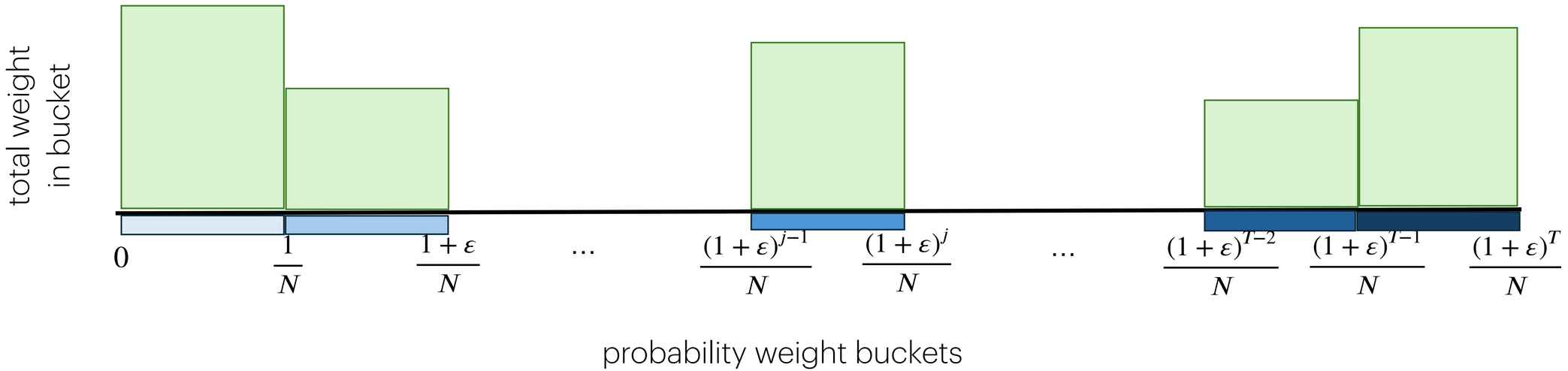
probability weights

Bucket Histograms

Learning a distribution's ε-bucket histogram suffices to learn any label-invariant property

[Batu-Fischer-Fortnow-Kumar-Rubinfeld-White-01, ..., Herman-Rothblum22, Herman-Rothblum23]

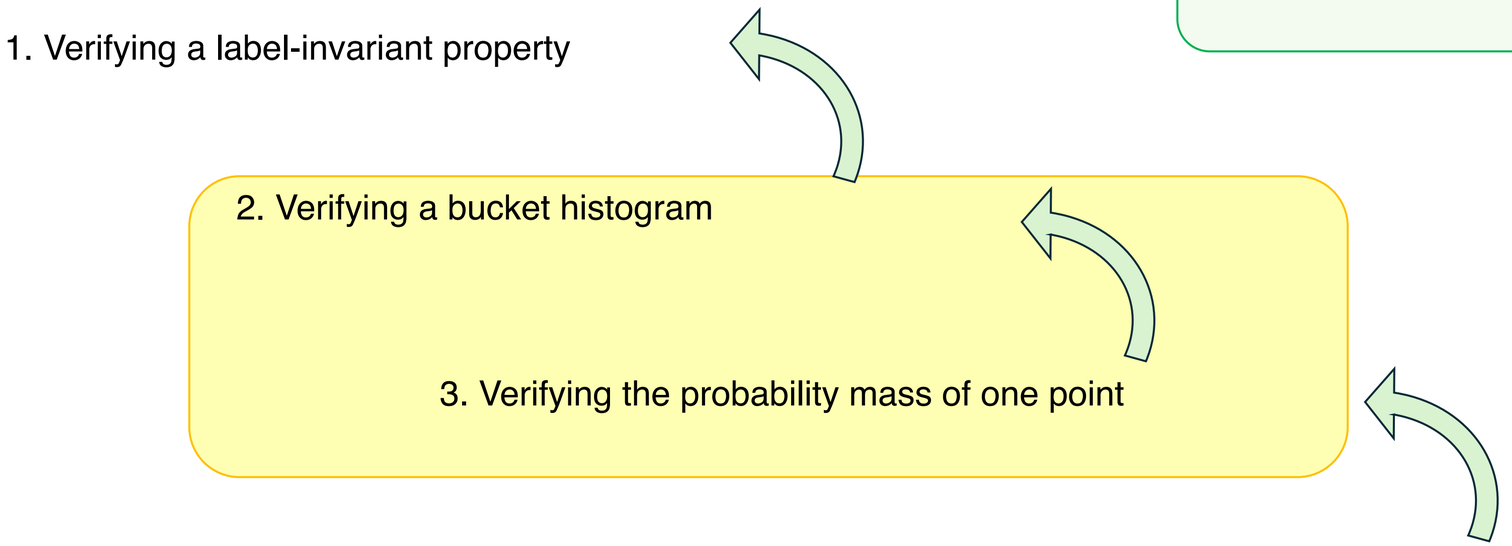
Only need to learn a histogram over $T \approx \frac{\log N}{\varepsilon}$



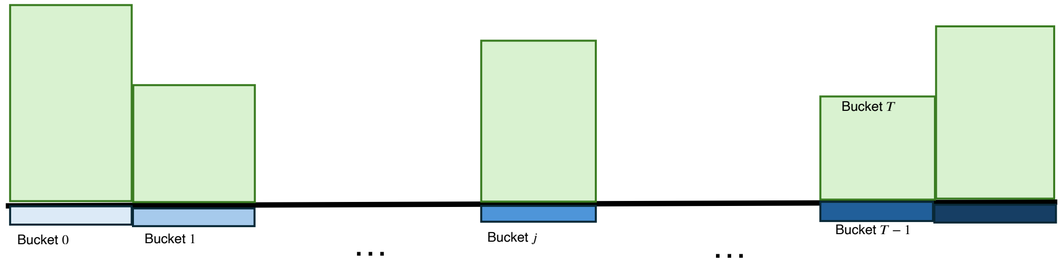
Proof Overview

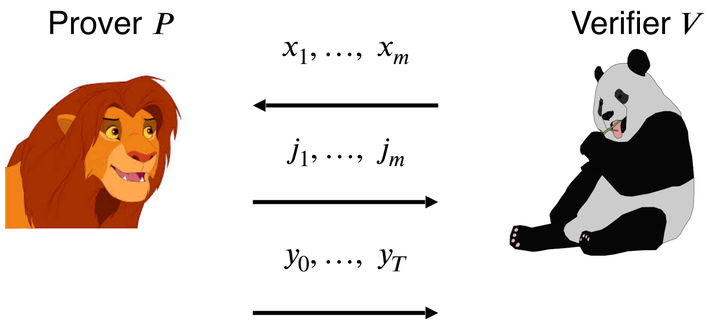
Every label-invariant property can be verified using

- $\mathrm{polylog}\, N$ samples,
- $\mathrm{polylog}\, N$ PCOND queries, and
- $\tilde{O}\left(\sqrt{N}\right)$ communication



Verifying A Bucket Histogram





1. Verifier samples $x_1, \ldots, x_m \sim D$ for some $m = \tilde{O}(T)$

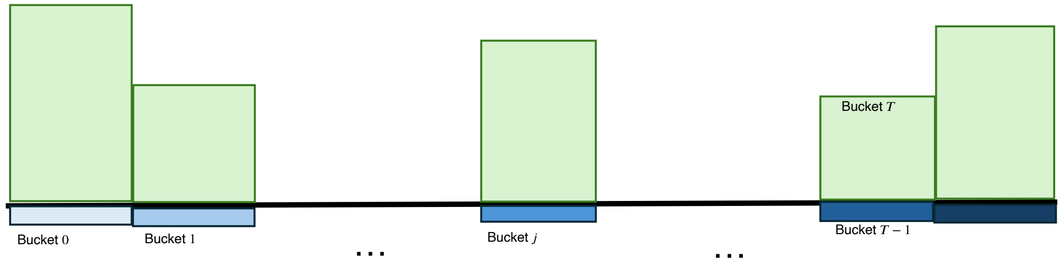
2. Prover sends back "tags" $j_1, \ldots, j_m$ claiming sample x_i is from bucket j_i

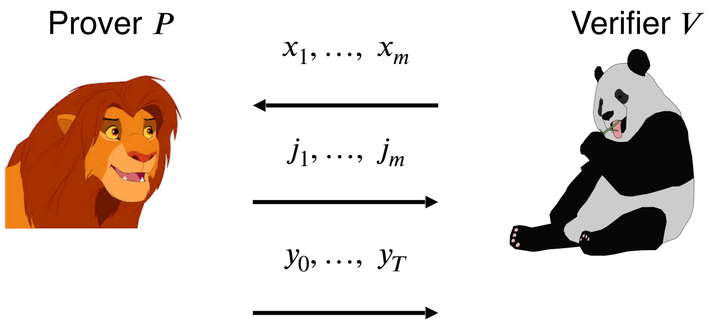
3. Prover sends "anchor points" $y_0, \ldots, y_T$ claiming that each
   a) y_j is in bucket j
   b) y_j has a "heavy neighbourhood"

We show that these anchor points exist w.h.p

4. Verifier uses $PCOND$ queries to compare pairs $(D[x_i], D[y_{j_i}])$ to confirm bucket tags

Verifying A Bucket Histogram





1. Verifier samples $x_1, \ldots, x_m \sim D$ for some $m = \tilde{O}(T)$

2. Prover sends back "tags" $j_1, \ldots, j_m$ claiming sample x_i is from bucket j_i

3. Prover sends "anchor points" $y_0, \ldots, y_T$ claiming that each
   a) y_j is in bucket j
   b) y_j has a "heavy neighbourhood"

   We show that these anchor points exist w.h.p

Prover could have LIED

4. Verifier uses $PCOND$ queries to compare pairs $(D[x_i], D[y_{j_i}])$ to confirm bucket tags

Proof Overview

Every label-invariant property can be verified using

- $\mathrm{polylog}\, N$ samples,
- $\mathrm{polylog}\, N$ PCOND queries, and
- $\tilde{O}\left(\sqrt{N}\right)$ communication

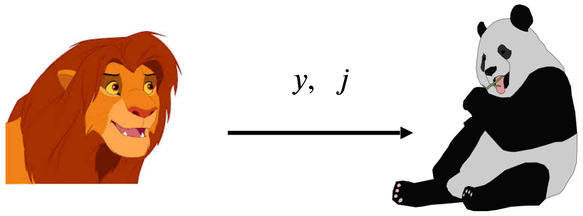
1. Verifying a label-invariant property

2. Verifying a bucket histogram

3. Verifying the probability mass of one point

4. Verifying the support size of a nearly-flat distribution

Verifying The Mass Of Anchor Points



This work: Can simulate samples from this distribution using PCOND

To verify prover's claims about y, suffices to check support size of $D|_{U(y)}$

Prover's claims:

a) y is in bucket j, i.e.,

$$\frac{(1+\varepsilon)^{j-1}}{N} \le D[y] \le \frac{(1+\varepsilon)^{j}}{N}$$

b) $D\left[U(y)\right] \ge \frac{1}{\text{polylog}N}$

where $U(y) = \{z \in [N] : D[z] \in (1 \pm \varepsilon)D[y]\}$

2. Verifier uses PCOND queries to estimate $D\left[U(y)\right]$ [Canonne-Ron Servedio15]

$$D\left[U(y)\right] \approx D[y] \cdot \left|U(y)\right|$$

$$p \approx D[y] \iff \left|U(y)\right| \approx \frac{D\left[U(y)\right]}{p}$$

Proof Overview

Every label-invariant property can be verified using

- $\mathrm{polylog}\, N$ samples,
- $\mathrm{polylog}\, N$ PCOND queries, and
- $\tilde{O}\left(\sqrt{N}\right)$ communication

1. Verifying a label-invariant property

2. Verifying a bucket histogram

3. Verifying the probability mass of one point

4. Verifying the support size of a nearly-flat distribution

Verifying Support Size

Assumption:

D is "almost flat": $\frac{1}{2} \leq \frac{D[x]}{D[y]} \leq 2$ for all x, y in support

Prover's claim: D is supported on A elements

Let S = true support of D

Prover's claim:
$A < \sqrt{N}$

Prover → Verifier:
Send claimed support $\tilde{S}$

Possible lie:
$|S| \gg |\tilde{S}|$

Verifier:
Sample from D and check it is in $\tilde{S}$

Possible lie:
$|S| \ll |\tilde{S}|$

Verifier:
Draw one point $x \sim D$
Sample uniformly from $\tilde{S}$ and compare mass to $D[x]$

Prover's claim:
$A \geq \sqrt{N}$

Possible lie:
$|S| \gg A$

Verifier → Prover:
Draw one point $x \sim D$, hide it amongst N/A samples from $[N]$
Prover → Verifier: Location of x

Possible lie:
$|S| \ll A$

Verifier → Prover:
Draw N/A samples from $[N]$
Prover → Verifier: Identify some x in S
Verifier: Draw one point from D and compare to x