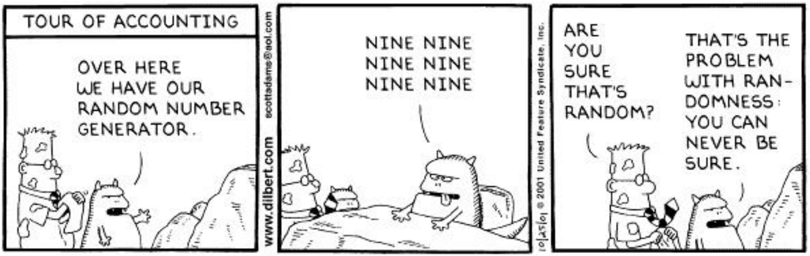
Generalised Uniformity Testing

Tuğkan Batu   Clément Canonne

Workshop on Data Summarization, University of Warwick
19 March 2018



Testing Distributions

Asking questions such as

- Are two distributions similar?
- Are two random variables independent?
- Is the distribution monotone?
- What is the entropy of the distribution?

Testing Distributions

Asking questions such as

- Are two distributions similar?
- Are two random variables independent?
- Is the distribution monotone?
- What is the entropy of the distribution?

Access to samples from the distribution, not explicit probabilities

Testing Distributions

Asking questions such as

- Are two distributions similar?
- Are two random variables independent?
- Is the distribution monotone?
- What is the entropy of the distribution?

Access to samples from the distribution, not explicit probabilities

No assumptions on the underlying distribution

Testing Distributions

Asking questions such as

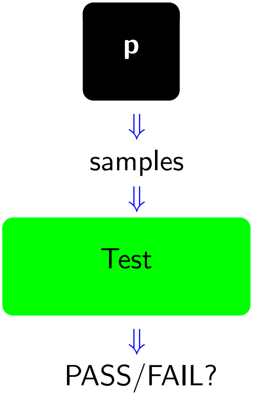
- Are two distributions similar?
- Are two random variables independent?
- Is the distribution monotone?
- What is the entropy of the distribution?

Access to samples from the distribution, not explicit probabilities

No assumptions on the underlying distribution

Focus: large discrete domains/alphabets

Testing Distributions



- $[n] = \{1, \ldots, n\}$ (typically known)
- $\mathbf{p}$: black-box distribution over $[n]$
  - generates i.i.d. samples
- $p_i = \Pr[\,\mathbf{p} \text{ outputs } i\,]$
- Error probability < 0.01
- Sample complexity in terms of n?

A Brief History of Testing Distributions

Many results in testing of discrete distributions over domain $[n]$: uniformity, identity, closeness, independence, monotonicity, log-concavity, juntas, MHR, PBD, SIIRV, histograms, ... [GR00,BFR+00,BFF+01, BKR04, Pan08, LRR11, VV14, ADK15, DKN15, BFR+10, CDVV14, Can16, DK16, DKS17,...]

A Brief History of Testing Distributions

Many results in testing of discrete distributions over domain $[n]$: uniformity, identity, closeness, independence, monotonicity, log-concavity, juntas, MHR, PBD, SIIRV, histograms, …[GR00,BFR+00,BFF+01, BKR04, Pan08, LRR11, VV14, ADK15, DKN15, BFR+10, CDVV14, Can16, DK16, DKS17,…]

We focus on
Uniformity.

Testing Uniformity

Lower bound (Impossibility):

$\Omega(\sqrt{n})$ samples are needed

- Consider $\mathbf{p} = U_{[n]}$ and $\mathbf{p}' = U_{[n/2]}$
- In $o(\sqrt{n})$ samples from $\mathbf{p}$ (or $\mathbf{p}'$), no repetitions (Birthday Problem)

Testing Uniformity

Lower bound (Impossibility):

$\Omega(\sqrt{n})$ samples are needed

- Consider $\mathbf{p} = U_{[n]}$ and $\mathbf{p}' = U_{[n/2]}$
- In $o(\sqrt{n})$ samples from $\mathbf{p}$ (or $\mathbf{p}'$), no repetitions (Birthday Problem)

Upper bound (Algorithm):

Techniques from [Goldreich Ron '00] extend to give a Uniformity Test with sample complexity $O(\sqrt{n}/\epsilon^4)$

- Estimate collision probability

Testing Uniformity

Lower bound (Impossibility):

$\Omega(\sqrt{n})$ samples are needed

- Consider $\mathbf{p} = U_{[n]}$ and $\mathbf{p}' = U_{[n/2]}$
- In $o(\sqrt{n})$ samples from $\mathbf{p}$ (or $\mathbf{p}'$), no repetitions (Birthday Problem)

Upper bound (Algorithm):

Techniques from [Goldreich Ron '00] extend to give a Uniformity Test with sample complexity $O(\sqrt{n}/\epsilon^4)$

- Estimate collision probability

[Paninski '08] shows sample complexity $\Theta(\sqrt{n}/\epsilon^2)$.

Shortcoming(s) of Known Results

- Domain size n must be given as input.

Shortcoming(s) of Known Results

- Domain size n must be given as input.
  - It may be unavailable.

Shortcoming(s) of Known Results

- Domain size n must be given as input.
  - It may be unavailable.
  - It may be irrelevant.

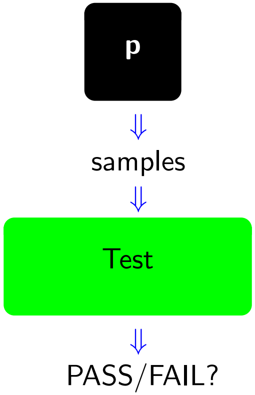
Shortcoming(s) of Known Results

- Domain size n must be given as input.
  - It may be unavailable.
  - It may be irrelevant.
- Non-adaptive algorithms that always match the worst case.

Shortcoming(s) of Known Results

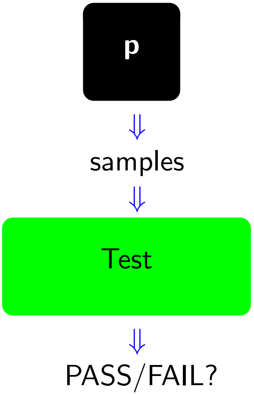
- Domain size n must be given as input.
  - It may be unavailable.
  - It may be irrelevant.
- Non-adaptive algorithms that always match the worst case.
- Usually not optimal for the given input distribution.

Testing Distributions Obliviously



- **p**: black-box distribution over unknown $S \subseteq \mathbb{N}$
  - generates i.i.d. samples
- $p_i = \Pr[\,\mathbf{p} \text{ outputs } i\,]$
- Error probability < 0.01
- Sample complexity in terms of some $f(\mathbf{p})$?

Testing Distributions Obliviously

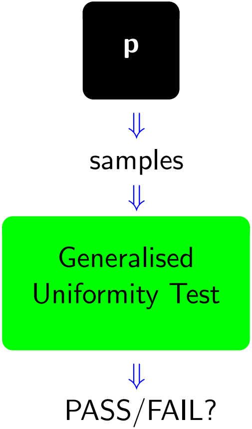


- $\mathbf{p}$: black-box distribution over unknown $S \subseteq \mathbb{N}$
  - generates i.i.d. samples
- $p_i = \Pr[\, \mathbf{p} \text{ outputs } i\,]$
- Error probability < 0.01
- Sample complexity in terms of some $f(\mathbf{p})$?

Questions:

- What should $f(\mathbf{p})$?
- How to detect when it has a large enough sample set?
- Optimal for each input distribution?

Generalised Uniformity Testing

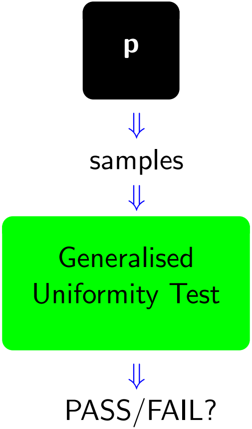


Goal:

- If $\mathbf{p} = U_S$ for some $S \subset \mathbb{N}$, then PASS
- If, $\forall S \subseteq \mathbb{N}$, $\Delta(\mathbf{p}, U_S) > \epsilon$, then FAIL

$\Delta(\cdot, \cdot)$: total variation distance

Generalised Uniformity Testing



Goal:

- If $\mathbf{p} = U_S$ for some $S \subset \mathbb{N}$, then PASS
- If, $\forall S \subseteq \mathbb{N}$, $\Delta(\mathbf{p}, U_S) > \epsilon$, then FAIL

  $\Delta(\cdot, \cdot)$: total variation distance

- How many samples are needed?
- How do we know when to stop?

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

$$\Pr[\underbrace{s_1 = s_2}_{\text{collision}}] = \sum_i p_i^2$$

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

$$\Pr[\underbrace{s_1 = s_2}_{\text{collision}}] = \sum_i p_i^2$$

Lemma ([Goldreich Ron 00])

Using $O(\sqrt{n})$ samples, we can estimate $\|\mathbf{p}\|_2^2$ very well.

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

$$\Pr[\underbrace{s_1 = s_2}_{\text{collision}}] = \sum_i p_i^2$$

Lemma ([Goldreich Ron 00])

Using $O(\sqrt{n})$ samples, we can estimate $\|\mathbf{p}\|_2^2$ very well.

m samples

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

$$\Pr[\underbrace{s_1 = s_2}_{\text{collision}}] = \sum_i p_i^2$$

Lemma ([Goldreich Ron 00])

Using $O(\sqrt{n})$ samples, we can estimate $\|\mathbf{p}\|_2^2$ very well.

m samples $\rightarrow$ $\binom{m}{2}$ pairs of samples

Testing Uniformity via Collision Probability

Definition

Collision probability of $\mathbf{p}$: $\sum_i p_i^2$ (a.k.a., $\|\mathbf{p}\|_2^2$)

$$\Pr[\underbrace{s_1 = s_2}_{\text{collision}}] = \sum_i p_i^2$$

Lemma ([Goldreich Ron 00])

Using $O(\sqrt{n})$ samples, we can estimate $\|\mathbf{p}\|_2^2$ very well.

m samples $\rightarrow$ $\binom{m}{2}$ pairs of samples $\rightarrow$ estimate $\|\mathbf{p}\|_2^2$ for $m \approx \sqrt{n}$

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.
  Idea: Wait until you see a collision.

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.
  More robust idea: Wait until you see k collisions.

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.
  More robust idea: Wait until you see k collisions.

Lemma

For any distribution $\mathbf{p}$, *we can estimate* $\|\mathbf{p}\|_2^2$ *within* $1 \pm \epsilon$ *using* $\Theta(\frac{1}{\epsilon^2 \cdot \|\mathbf{p}\|_2})$ *samples.*

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.
  More robust idea: Wait until you see k collisions.

Lemma

For any distribution $\mathbf{p}$*, we can estimate* $\|\mathbf{p}\|_2^2$ *within* $1 \pm \epsilon$ *using* $\Theta(\frac{1}{\epsilon^2 \cdot \|\mathbf{p}\|_2})$ *samples.*

Tight in terms of $\|\mathbf{p}\|_2$.

Generalised Uniformity Testing

Estimating $\|\mathbf{p}\|_2^2$

- For a uniform $\mathbf{p}$, $\frac{1}{\|\mathbf{p}\|_2^2}$ is the support size.
- Maybe we can still use $\|\mathbf{p}\|_2^2$.
  More robust idea: Wait until you see k collisions.

Lemma

For any distribution $\mathbf{p}$, we can estimate $\|\mathbf{p}\|_2^2$ within $1 \pm \epsilon$ using $\Theta(\frac{1}{\epsilon^2 \cdot \|\mathbf{p}\|_2})$ samples.

Tight in terms of $\|\mathbf{p}\|_2$.

Lemma

Estimating $\|p\|_2^2$ requires $\Omega(\frac{1}{\|\mathbf{p}\|_2})$ samples.

We got $\|\mathbf{p}\|_2^2$! What do we do?

We got $\|\mathbf{p}\|_2^2$! What do we do?

- Two distributions can have the same norm, but different profiles.

We got $\|\mathbf{p}\|_2^2$! What do we do?

- Two distributions can have the same norm, but different profiles.
- We need to observe 3-way collisions:

$$\Pr[s_1 = s_2 = s_3] = \sum_i p_i^3$$

We got $\|\mathbf{p}\|_2^2$! What do we do?

- Two distributions can have the same norm, but different profiles.
- We need to observe 3-way collisions:

$$\Pr[s_1 = s_2 = s_3] = \sum_i p_i^3$$

We got $\|\mathbf{p}\|_2^2$! What do we do?

- Two distributions can have the same norm, but different profiles.
- We need to observe 3-way collisions:

$$\Pr[s_1 = s_2 = s_3] = \sum_i p_i^3$$

Observation: For a fixed value for $\|\mathbf{p}\|_2$, the uniform distribution on $\frac{1}{\|\mathbf{p}\|_2^2}$ will generate the fewest 3-way collisions.

We got $\|\mathbf{p}\|_2^2$! What do we do?

- Two distributions can have the same norm, but different profiles.
- We need to observe 3-way collisions:

$$\Pr[s_1 = s_2 = s_3] = \sum_i p_i^3$$

Observation: For a fixed value for $\|\mathbf{p}\|_2$, the uniform distribution on $\frac{1}{\|\mathbf{p}\|_2^2}$ will generate the fewest 3-way collisions.

Lemma

Let $\mathbf{p}$ *be a distribution over* $\mathbb{N}$ *and* $N \in \mathbb{N}$ *such that*

$$\frac{1-\epsilon}{N} \leq \|\mathbf{p}\|_2^2 \leq \frac{1+\epsilon}{N} \qquad \text{and} \qquad \|\mathbf{p}\|_3^3 \leq \frac{1+\delta}{N^2},$$

for some $0 < \epsilon, \delta < 0.04$*. Then, the* ℓ_1 *distance of* $\mathbf{p}$ *to any uniform distribution* $\mathbf{q}$ *can be upper bounded as*

$$\Delta(\mathbf{p}, \mathbf{q}) \leq 9\sqrt[3]{\delta + 3\epsilon}.$$

Putting It Altogether

Algorithm $\textsc{Test-Uniformity}(\mathbf{p}, \epsilon)$

$\delta \leftarrow O(\epsilon^3)$, $k \leftarrow \lceil \epsilon^{-18} \rceil$

$N \leftarrow 1/\textsc{Estimate-}\ell_2\textsc{-norm}(\mathbf{p}, \delta)$

Keep taking samples from $\mathbf{p}$ until k 3-way collisions are observed or $M = \sqrt[3]{3(1-4\delta)k}N^{2/3}$ samples are taken

if more than k 3-way collisions are observed **then**

  return REJECT

else

  return ACCEPT

Putting It Altogether

Algorithm TEST-UNIFORMITY($\mathbf{p}, \epsilon$)
  $\delta \leftarrow O(\epsilon^3)$, $k \leftarrow \lceil \epsilon^{-18} \rceil$
  $N \leftarrow 1/\text{ESTIMATE-}\ell_2\text{-NORM}(\mathbf{p}, \delta)$
  Keep taking samples from $\mathbf{p}$ until k 3-way collisions are observed or $M = \sqrt[3]{3(1-4\delta)k}N^{2/3}$ samples are taken
  if more than k 3-way collisions are observed **then**
    return REJECT
  else
    return ACCEPT

Theorem

The test above, with probability at least $3/4$*, accepts a uniform distribution and rejects a distribution* ϵ*-far from any uniform distribution. The expected sample complexity is* $\Theta(\frac{1}{\epsilon^6 \cdot ||\mathbf{p}||_3})$.

Putting It Altogether

Algorithm $\textsc{Test-Uniformity}(\mathbf{p}, \epsilon)$
    $\delta \leftarrow O(\epsilon^3)$, $k \leftarrow \lceil \epsilon^{-18} \rceil$
    $N \leftarrow 1/\textsc{Estimate-}\ell_2\textsc{-norm}(\mathbf{p}, \delta)$
    Keep taking samples from $\mathbf{p}$ until k 3-way collisions are observed or $M = \sqrt[3]{3(1-4\delta)k}N^{2/3}$ samples are taken
    if more than k 3-way collisions are observed **then**
        return REJECT
    else
        return ACCEPT

Theorem

The test above, with probability at least $3/4$, accepts a uniform distribution and rejects a distribution ϵ-far from any uniform distribution. The expected sample complexity is $\Theta(\frac{1}{\epsilon^6 \cdot ||\mathbf{p}||_3})$.

Essentially tight. We certainly need $\Omega(\frac{1}{||\mathbf{p}||_3})$.

Instance-specific Lower Bound

Theorem

For any fixed non-uniform distribution $\mathbf{q}$*, distinguishing between*
(i) $\mathbf{p} = \mathbf{q}$ *(up to a permutation) and*
(ii) uniform $\mathbf{p}$
requires $\Omega(1/||\mathbf{p}||_3)$ *samples from* $\mathbf{p}$*.*

Instance-specific Lower Bound

Theorem

For any fixed non-uniform distribution $\mathbf{q}$*, distinguishing between*
(i) $\mathbf{p} = \mathbf{q}$ *(up to a permutation) and*
(ii) uniform $\mathbf{p}$
requires $\Omega(1/||\mathbf{p}||_3)$ *samples from* $\mathbf{p}$*.*

- Instance-specific **not** worst-case

Instance-specific Lower Bound

Theorem

For any fixed non-uniform distribution $\mathbf{q}$*, distinguishing between*
(i) $\mathbf{p} = \mathbf{q}$ *(up to a permutation) and*
(ii) uniform $\mathbf{p}$
requires $\Omega(1/||\mathbf{p}||_3)$ *samples from* $\mathbf{p}$*.*

- Instance-specific **not** worst-case
- Proof uses Wishful Thinking Theorem of [Valiant11].

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound
- Extensions to other distribution testing problems

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound
- Extensions to other distribution testing problems
- Instance-specific lower bounds

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound
- Extensions to other distribution testing problems
- Instance-specific lower bounds
- Other notions of distance between distributions

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound
- Extensions to other distribution testing problems
- Instance-specific lower bounds
- Other notions of distance between distributions
  - Earth-mover distance?

Remarks

- Follow-up work of Diakonikolas, Kane, and Stewart 17:
  - improve upper bound for ϵ dependence;
  - complement it with (worst-case) matching lower bound
- Extensions to other distribution testing problems
- Instance-specific lower bounds
- Other notions of distance between distributions
  - Earth-mover distance?
  - Domain with a metric?

Thank You!