Distributions Property Testing

V. Nikishkin

Laboratory for Foundations of Computer Science
School of Informatics
University of Edinburgh

September 2015

Table of Contents

What is property testing?

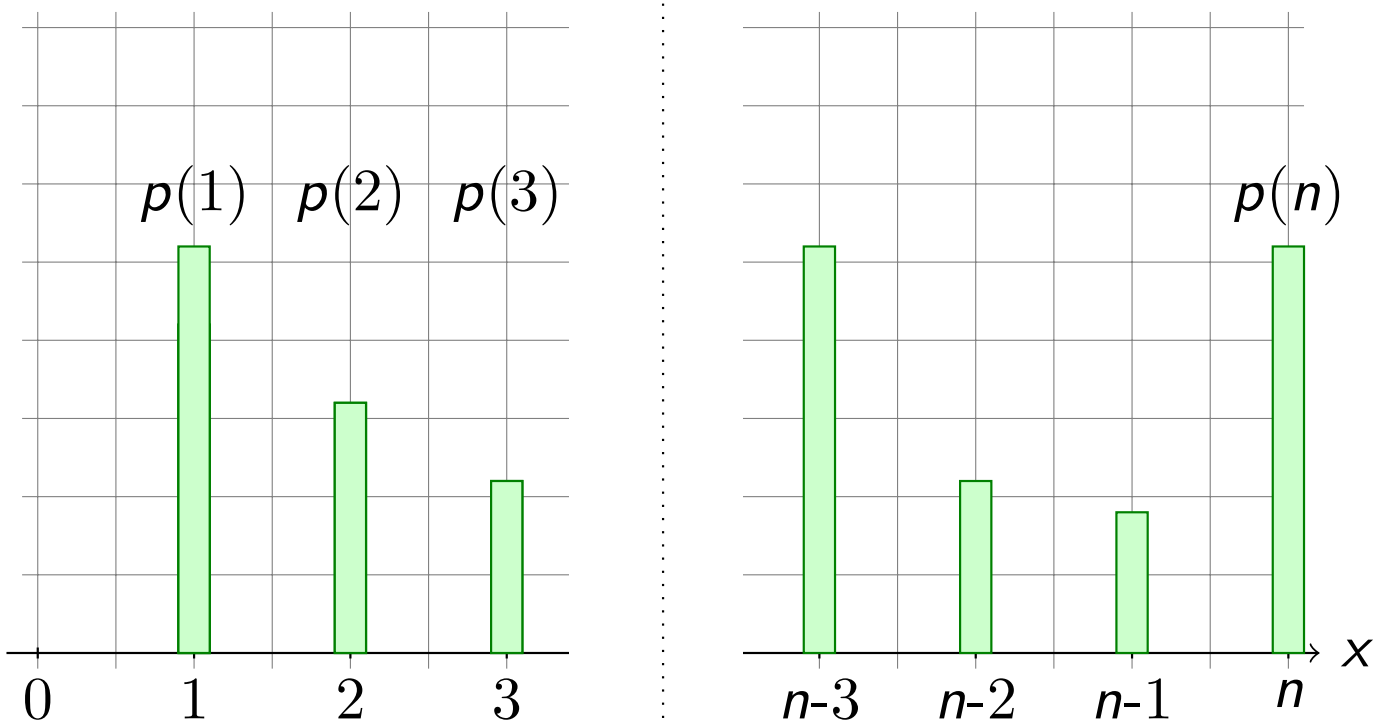
Statistical hypothesis testing applied to combinatorial/probailistic objects.

H_0: Object A has some property $\mathcal{B}$ ($A \in \mathcal{B}$)

H_1: Object A is far from some property $\mathcal{B}$ ($A \notin \mathcal{B}$)

Distributions

Discrete probability distributions over $[n] = \{1, 2, ..., n\}$.

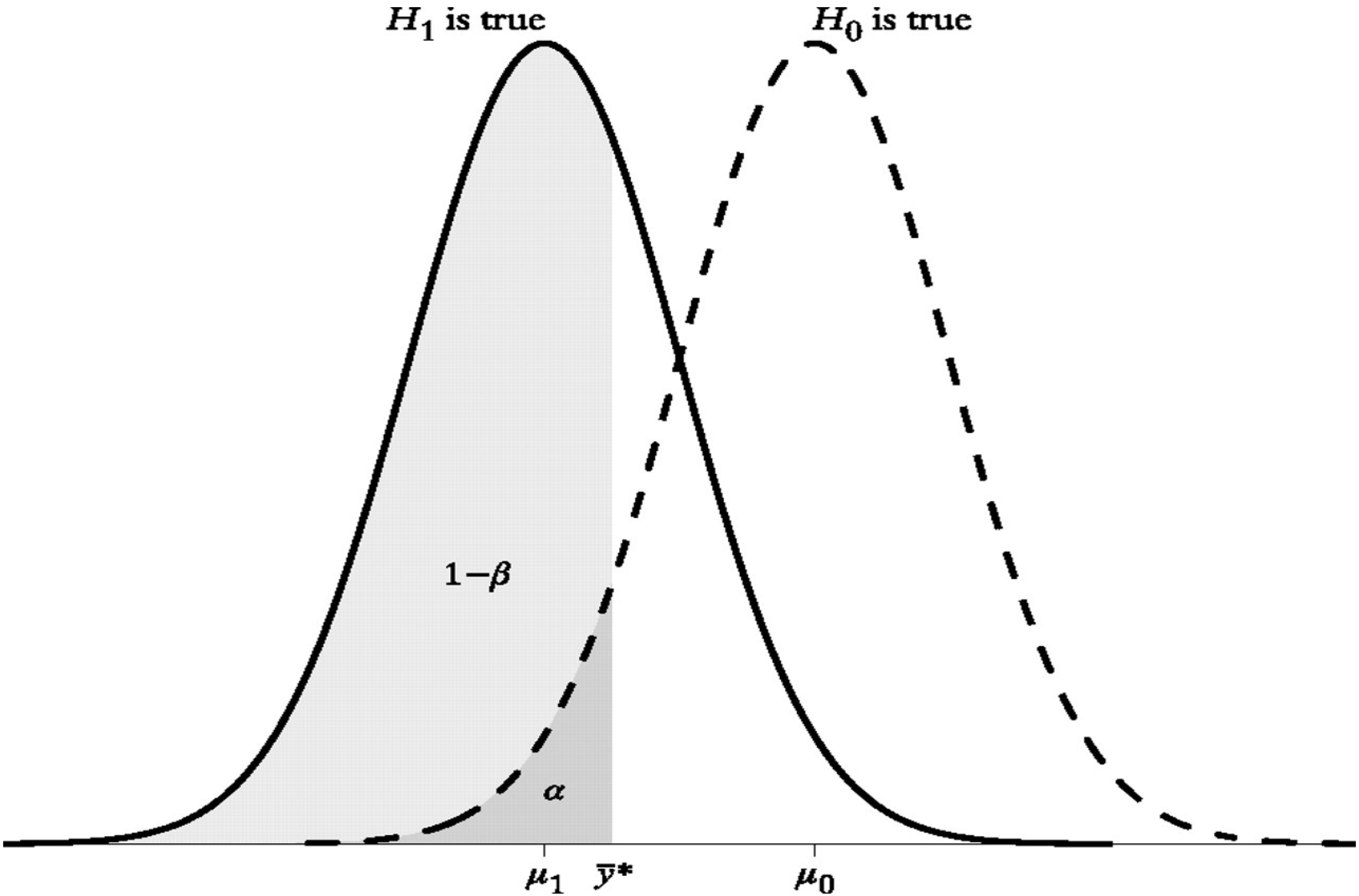


L_1 and L_2 distances between distributions are defines as:

$$L_1 : \quad \|p - q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

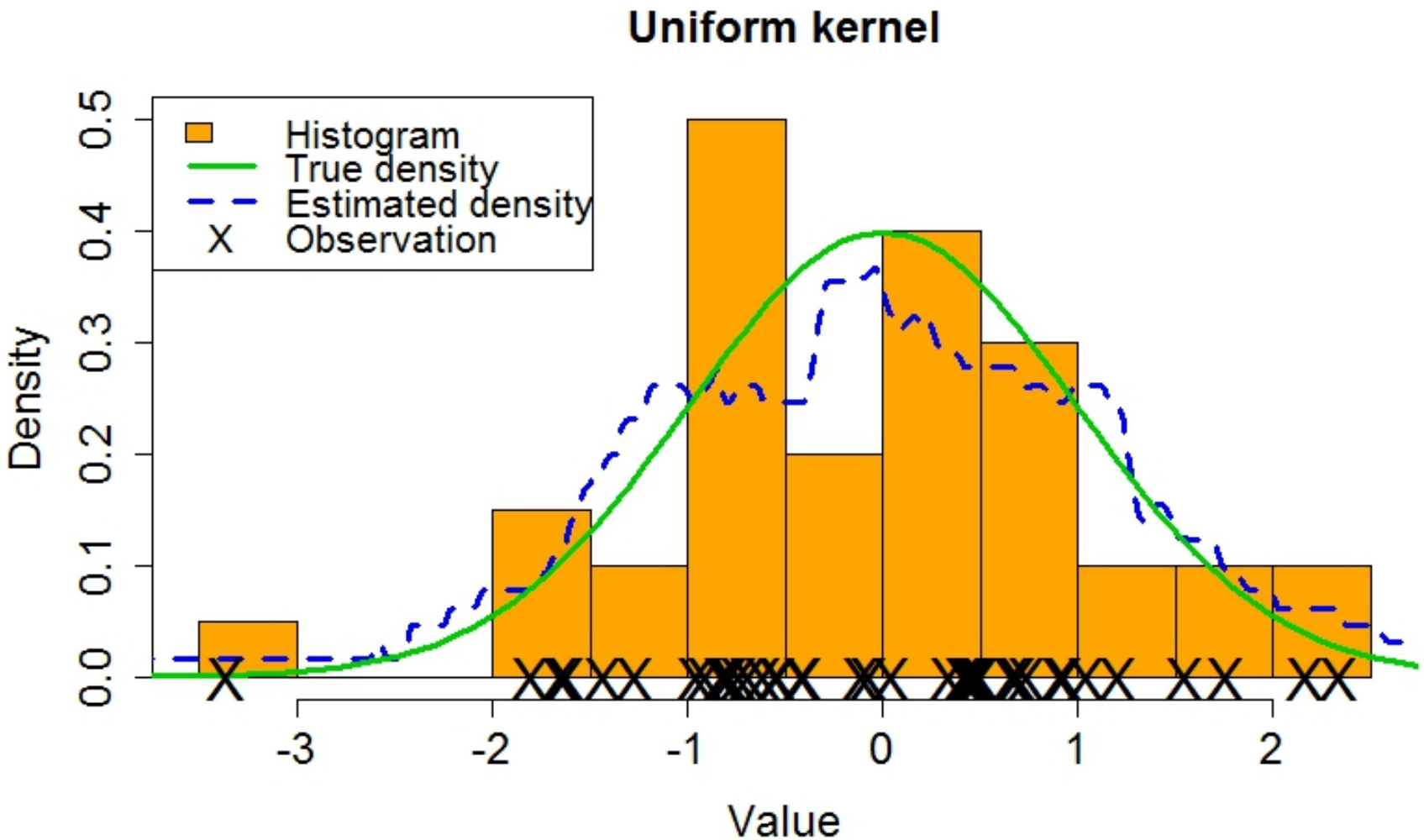
$$L_2 : \quad \|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Distribution of a criterion function $f(x)$



Easiest approach

Learn the distribution. $f(x) = I\{x \in \mathcal{B}\}$

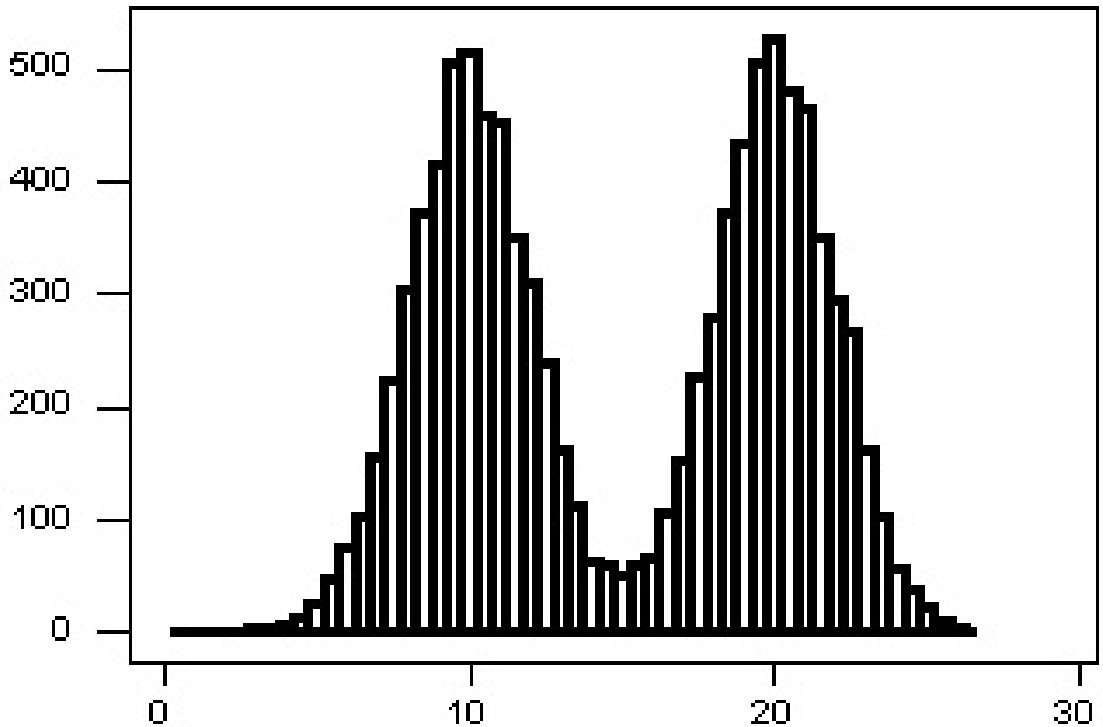


Why does it fail?

Sample size is too big.

$$C = \frac{n}{\varepsilon^2}$$

Shape restriction (Saving samples)



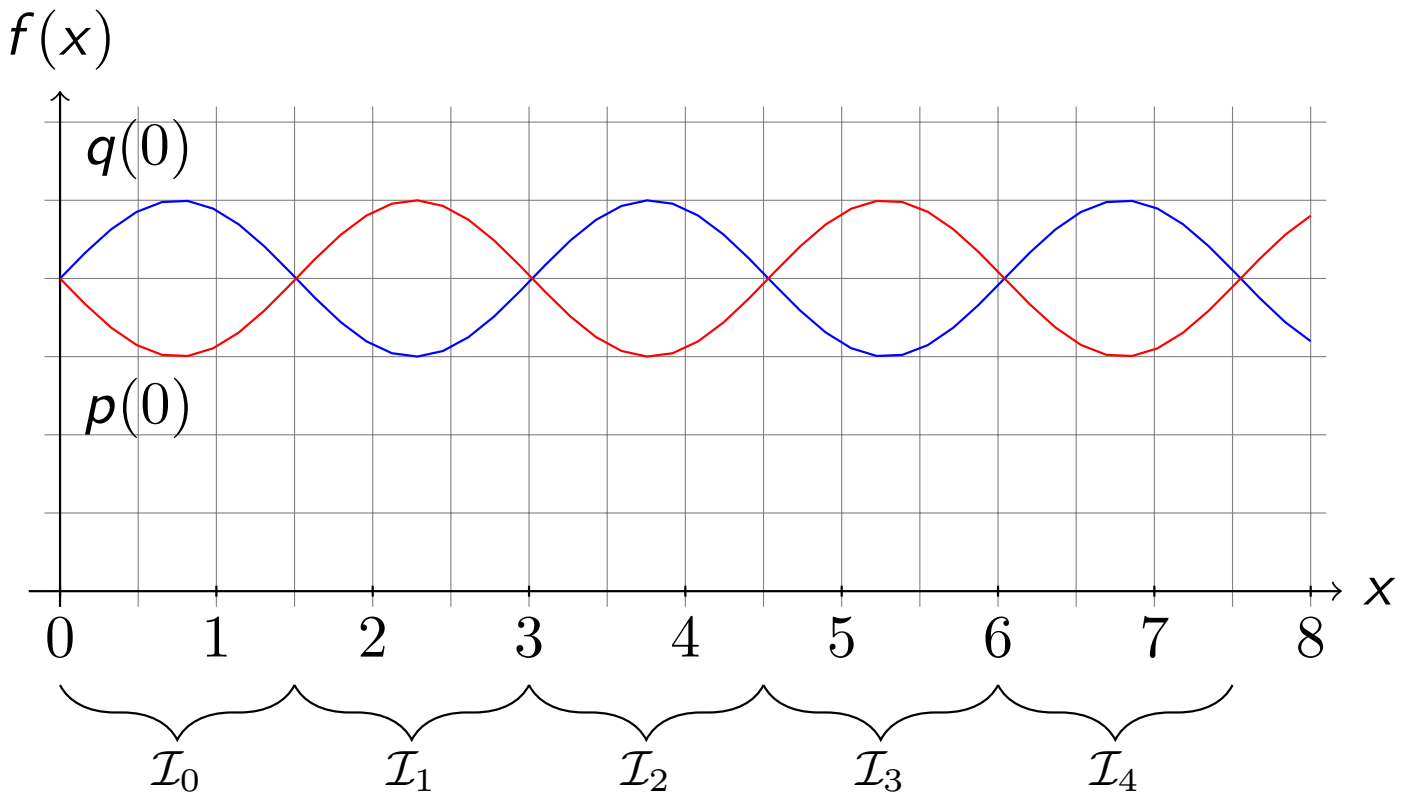
An example of a structured distribution.

A_k-distance

$$\|p - q\| \stackrel{\text{def}}{=} \max_{\mathcal{I}=(I_i)_{i=1}^k \in \mathcal{J}_k} \sum_{i=1}^{k} |p(I_i) - q(I_i)| = \max_{\mathcal{J}_k} \|p_r^{\mathcal{I}} - q_r^{\mathcal{I}}\|_{L_1}$$

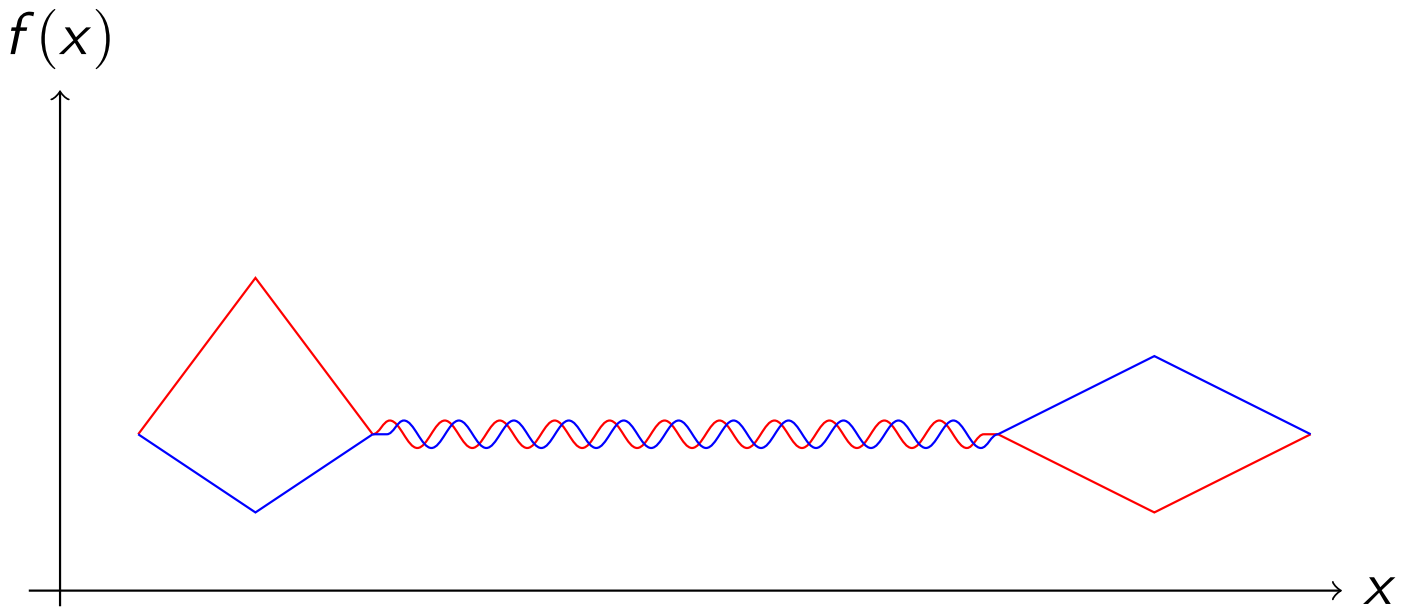
Where $\mathcal{J}_k$ denotes a set of partitions of $[1 \dots n]$ into k intervals.

A_k distance



Two distributions having "essentially" 5 crossings need $k = 5$.

Intuition



4 essential crossings

New results

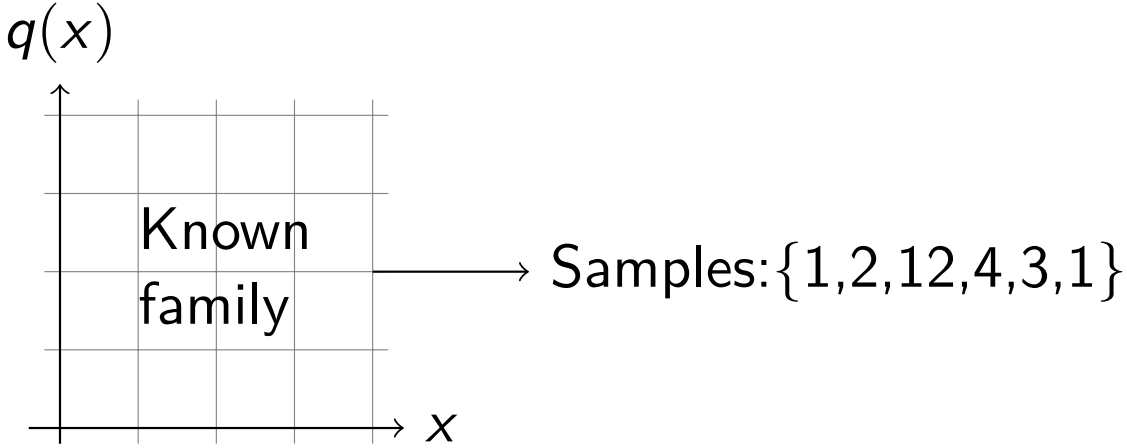
- Testing shape-restricted uniformity (A brief reminder about the last year work) [SODA2015]
- Testing shape-restricted identity [FOCS2015]
- Lower bound for shape-restricted identity [FOCS2015]

Table of Contents

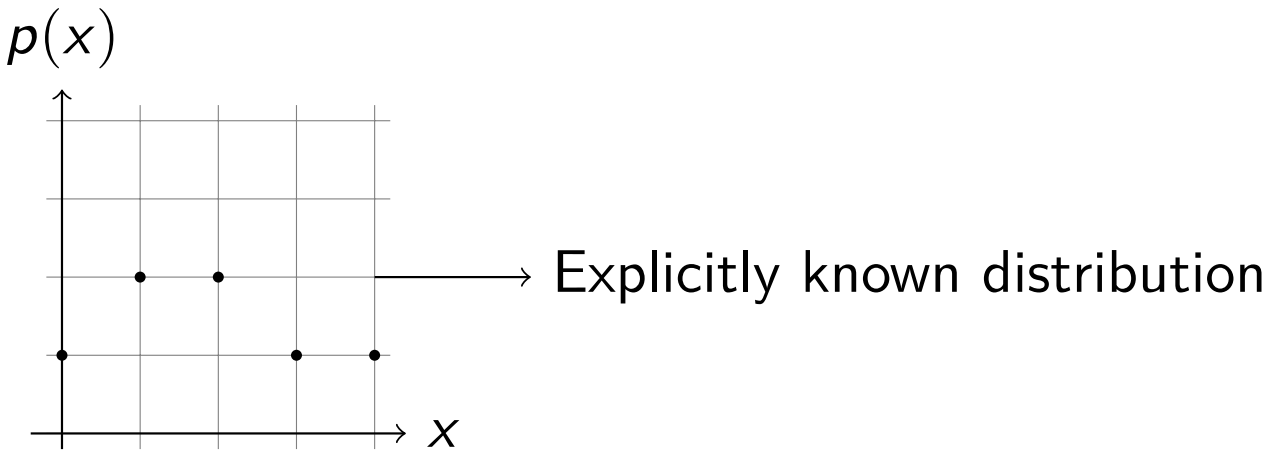
Problem definition

Distinguish between the cases $p = q$ and $L_1(p, q) \geqslant \varepsilon$.

- q :



- p :



Existing result is $\frac{\sqrt{n}}{\varepsilon^2}$.

Property testing

Problem:

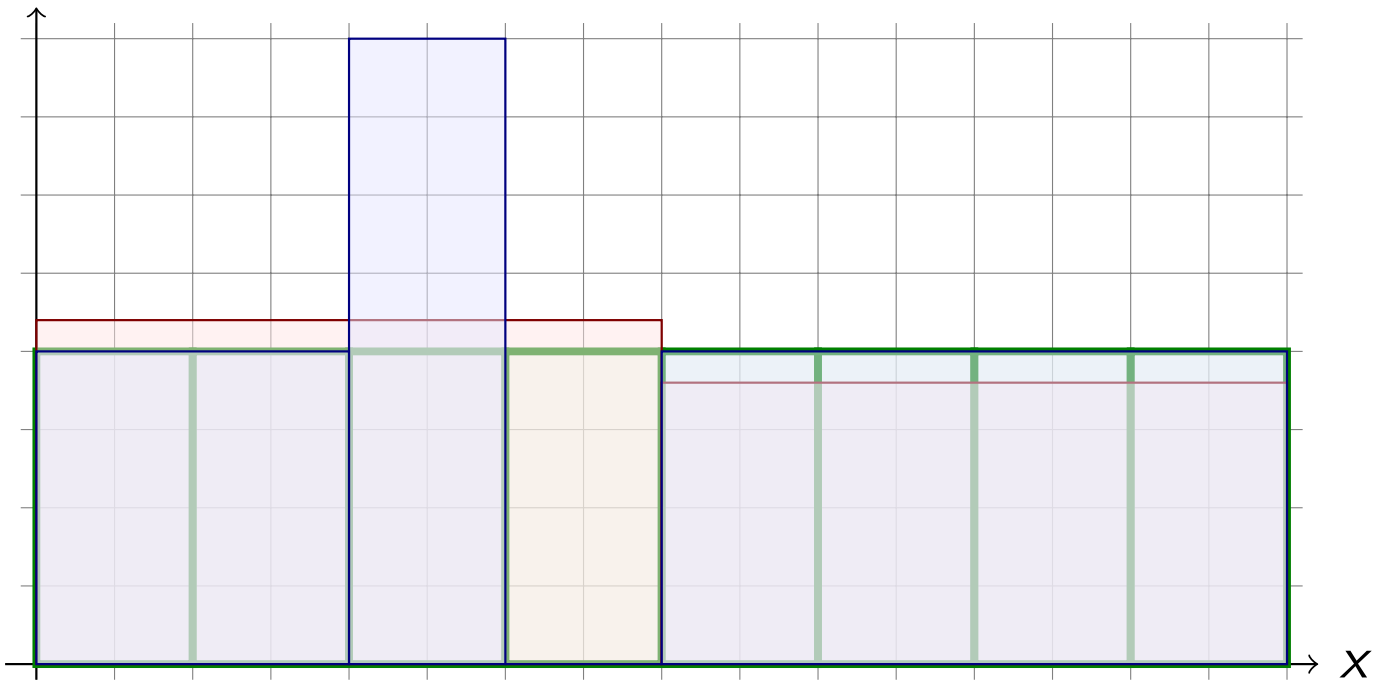
Identity testing. Distributions: p (oracle) and q (given), distance ε, error δ. TESTER$(p, q, \varepsilon, \delta) =$ "SAME" if $p \equiv q$, "FAR" if $dist(p, q) \geqslant \varepsilon$.

Distance metrics may be different: L_1, L_2, L_∞.

Algorithm

1. "Stretch the distribution"
2. "Partition the domain into many intervals"
3. "Test L_2 uniformity"
4. "Refine partition, decrease sensitivity of the tester"
5. "Go to step 3"

L_2 to L_1 testing (Tall vs. wide)



Subroutine: L_2 tester

The L_2 tester we use as a subroutine belongs to a family of "Chi-squared-like" testers.

$$\sum_{i=1}^{n} (X_i - m/n)^2 - X_i \geqslant 4m/\sqrt{n}$$

Applications of Main Result

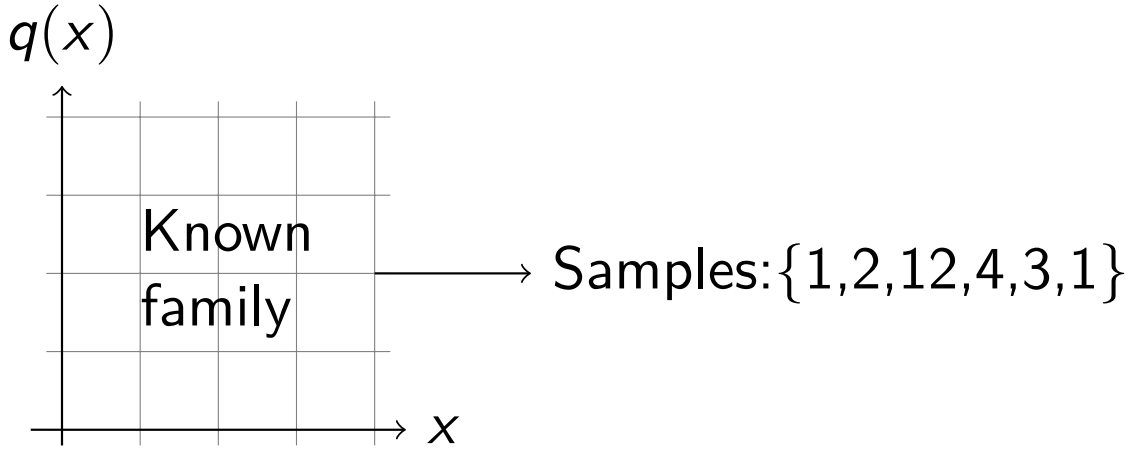
Class of Distributions over $[n]$	**Our upper bound**	**Previous work**	**k**
t-piecewise constant	$O(\sqrt{t}/\varepsilon^2)$	$O(t/\varepsilon^2)$ [CDSS14] (learn)	$O(t)$
t-piecewise degree-d polynomial	$O\left(\sqrt{t(d+1)}/\varepsilon^2\right)$	$O\left(t(d+1)/\varepsilon^2\right)$ [CDSS14] (learn)	$O(t(d+1))$
log-concave	$O(1/\varepsilon^{9/4})$	$\widetilde{O}(1/\varepsilon^{5/2})$ [CDSS14] (learn)	$O(1/\varepsilon^{1/2})$
ℓ-mixture of log-concave	$\sqrt{\ell} \cdot O(1/\varepsilon^{9/4})$	$\widetilde{O}(\ell/\varepsilon^{5/2})$ [CDSS14] (learn)	$O(\ell \cdot 1/\varepsilon^{1/2})$
t-modal	$O(\sqrt{t \log(n)}/\varepsilon^{5/2})$	$O\left(\sqrt{t \log(n)}/\varepsilon^3 + t^2/\varepsilon^4\right)$ [DDSVV13]	$O(t \log(n) \cdot 1/\varepsilon)$
ℓ-mixture of t-modal	$O(\sqrt{\ell t \log(n)}/\varepsilon^{5/2})$	$O\left(\sqrt{\ell t \log(n)}/\varepsilon^3 + \ell^2 t^2/\varepsilon^4\right)$ [DDSVV13]	$O(\ell t \log(n) \cdot 1/\varepsilon)$
monotone hazard rate (MHR)	$O(\sqrt{\log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14] (learn)	$O(\log(n/\varepsilon) \cdot 1/\varepsilon)$
ℓ-mixture of MHR	$O(\sqrt{\ell \log(n/\varepsilon)}/\varepsilon^{5/2})$	$O(\ell \log(n/\varepsilon)/\varepsilon^3)$ [CDSS14] (learn)	$O(\ell \log(n/\varepsilon) \cdot 1/\varepsilon)$

Table of Contents

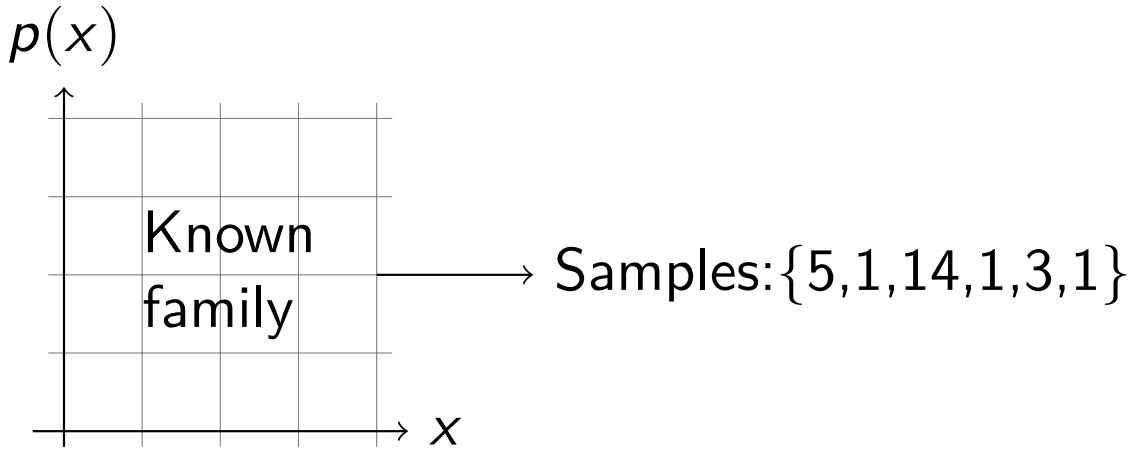
Problem definition

Distinguish between the cases $p = q$ and $L_1(p, q) \geqslant \varepsilon$.

- q :



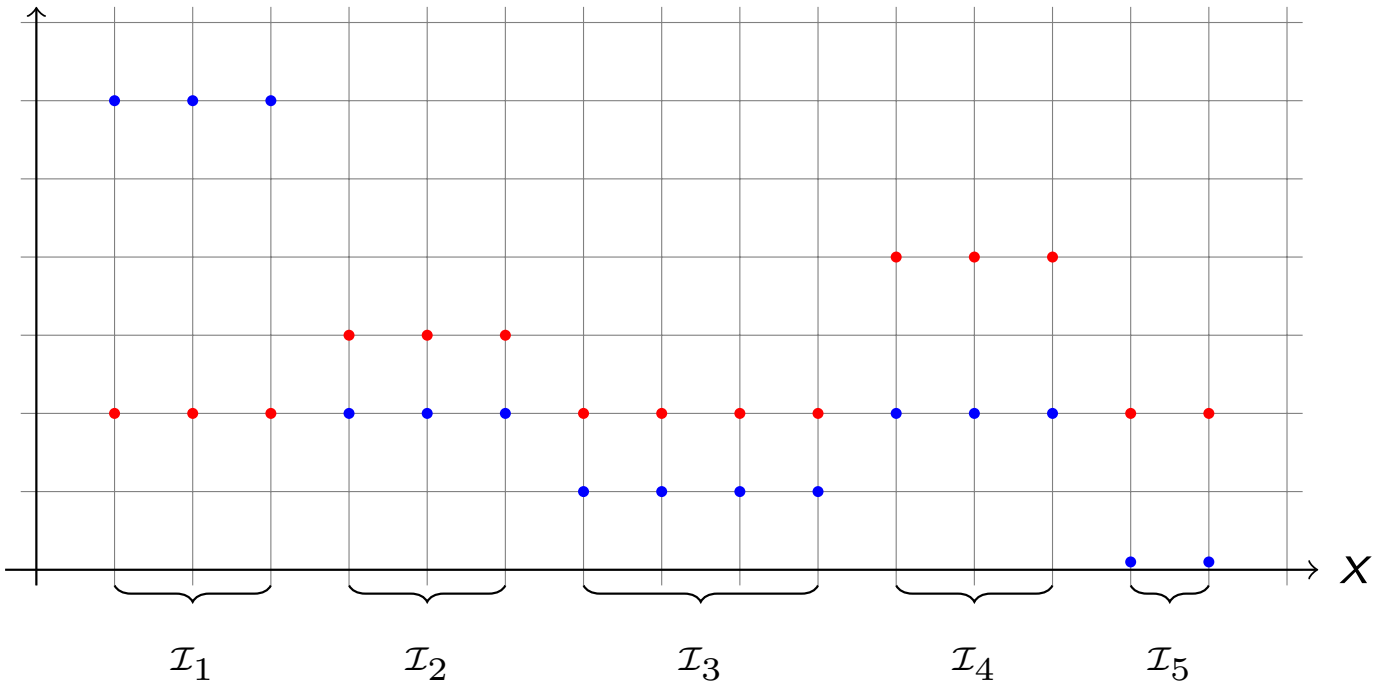
- p :



We want something like $O(k^{2/3})$ analogous to [CDVV] for L_1.

- CASE1: $p = q$
- CASE2: $\rho_{A_k}(p, q) \gg \varepsilon$

A_k is identical to L_1 for structured (e.g. k-flat) distributions.



But we don't know where $\mathcal{I}_{1\ldots k}$ are!
And we can't approximate, as it requires $O(k)$ samples.

Results

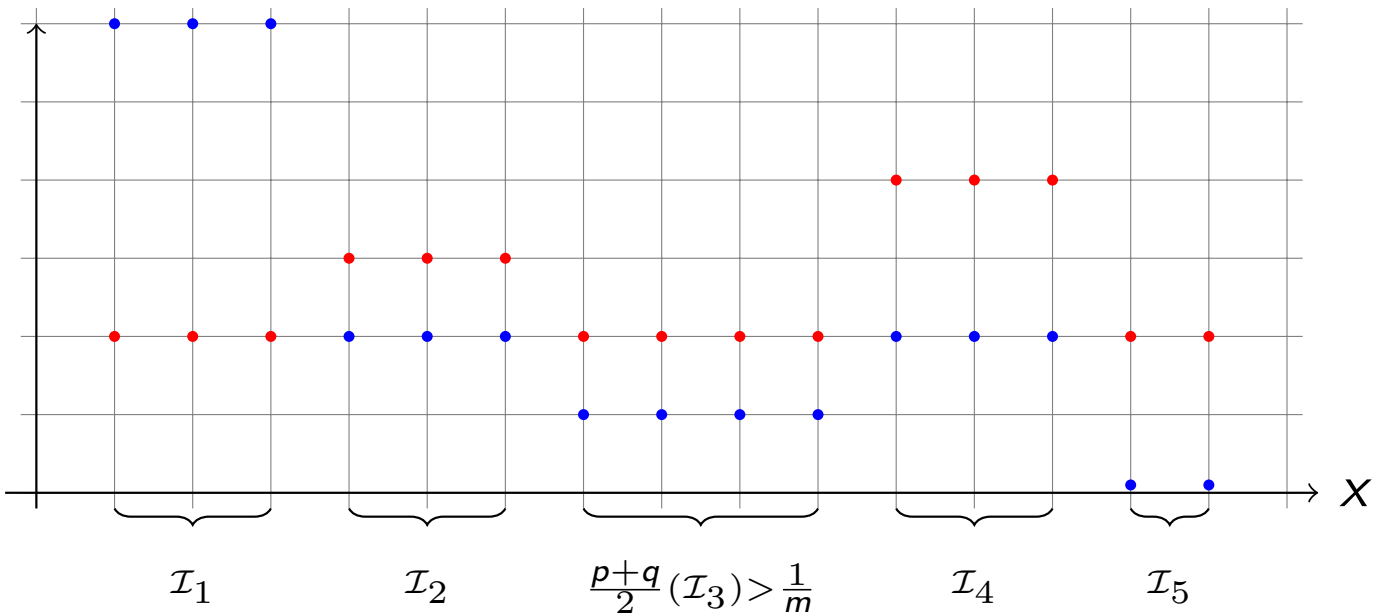
We did solve the problem.

Family of Distributions	**Our upper bound**	**Previous work**
t-piecewise constant	$O(\max\{t^{4/5}/\varepsilon^{6/5}, t^{1/2}/\varepsilon^2\})$	$O(t/\varepsilon^2)$ [CDSS14]
t-piecewise degree-d polynomial	$O(\max\{(t(d+1))^{4/5}/\varepsilon^{6/5}, (t(d+1))^{1/2}/\varepsilon^2\})$	$O\left(t(d+1)/\varepsilon^2\right)$ [CDSS14]
log-concave	$O(1/\varepsilon^{9/4})$	$O(1/\varepsilon^{5/2})$ [CDSS14, ADLS15]
k-mixture of log-concave	$O(\max\{k^{4/5}/\varepsilon^{8/5}, k^{1/2}/\varepsilon^{9/4}\})$	$O(k/\varepsilon^{5/2})$ [CDSS14, ADLS15]
t-modal over $[n]$	$O(\max\{(t\log n)^{4/5}/\varepsilon^2, (t\log n)^{1/2}/\varepsilon^{5/2}\})$	$O((t\log(n))^{2/3}/\varepsilon^{8/3} + t^2/\varepsilon^4)$ [DDSVV13]
monotone hazard rate (MHR) over $[n]$	$O(\max\{\log(n/\varepsilon)^{4/5}/\varepsilon^2, \log(n/\varepsilon)^{1/2}/\varepsilon^{5/2}\})$	$O(\log(n/\varepsilon)/\varepsilon^3)$ [CDSS14, ADLS15]

We need a better lower bound. (Next part of the talk.)

Big and small ε

Cases $\varepsilon > \frac{1}{k^{1/6}}$ and $\varepsilon < \frac{1}{k^{1/6}}$ differ significantly. The first tester fails if most of the discrepancy falls on "heavy" intervals.

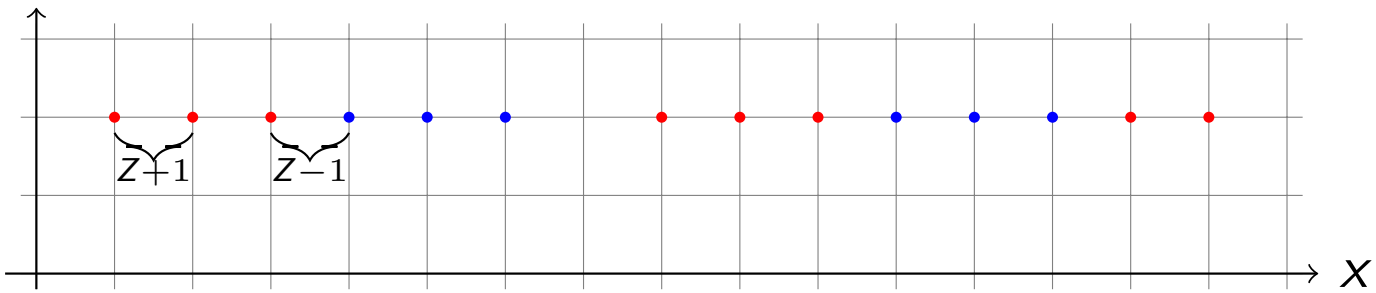


The algorithm TESTER1

Solution for big $\varepsilon(> \frac{1}{k^{1/6}})$: (The algorithm TESTER1)

1. Sample $m = O(\frac{k^{4/5}}{\varepsilon^{6/5}})$ samples from both distributions.
2. Sort the samples.
3. Compute
   $Z = \#(\text{successive samples from the same distribution}) -$
   $\#(\text{successive samples from different distributions})$
4. If $Z > 3\sqrt{m}$ return "DIFFERENT"

How does such a simple algorithm work?



Completeness

- We use Poissonization, i.e. sample $Poi(m)$ samples.
- Clearly, $E[Z] = 0$, $Var[Z] = 2m - 1$.
- By Chebyshev, $Pr\{\|Z\| < 3\sqrt{m}\} > 7/9$.

Soundness Preliminaries

We want to use Chebyshev, prove $E[Z] \gg \sqrt{Var[Z]} + \sqrt{m}$

$f(t) \stackrel{\text{def}}{=} Pr[\text{largest sample in S that is at most t was drawn from p}] -$
$-Pr[\text{largest sample in S that is at most t was drawn from q}]$

W.l.o.g. $p, q \in C[0,1]$ $\|p\|, \|q\| < 2$ (By using monotonous transformation.)

Soundness Expectation

1. Claim 1: $E[Z] = \int_0^1 f(t)(p(t) - q(t))dt$
2. Claim 2: $E[Z] = \int_0^1 f^2(t)(p(t) + q(t))dt + f^2(1)/2$
3. SubClaim 2.1: $f'(t) = m\Big(p(t) - q(t) - f(t)\Big(p(t) + q(t)\Big)\Big)$

For every interval $I \in [0,1]$ s.t. $\|p(I) - q(I)\| = \delta$ and $(p+q)(I) < 1/m$:

1. Claim 3: $\exists x \in I \ s.t. \|f(x)\| > 2m\delta/3$
2. Claim 4: $\int_I f^2(t)(p(t) + q(t))dt = \Omega(m^2\delta^3)$

We can partition $[0,1]$ into $3k$ intervals so that $p(I), q(I) < \frac{1}{k}$ and apply Claim 4:

$$E[Z] = \Omega(m\sum_{i=1}^{3k} m^2\delta_i^3) = \Omega(\frac{m^3\sum_{i=1}^{3k}\delta_i}{3k^2}) = \Omega(\frac{m^3\varepsilon^3}{k^2}) = \Omega(C^{5/2}\sqrt{m})$$

Soundness Variance

$$Var[Z] = O(m)$$

1. Partition $[0, 1]$ into m intervals $I_1 \ldots I_m$ so that $\frac{p+q}{2}(I_i) = 2/m$
2. $X_i \stackrel{\text{def}}{=}$ "contribution to Z coming from I_i"
3. $Var[Z] = \sum_{i,j} Cov[X_i, X_j]$
4. 4.1 $Cov[X_i, X_i] = Var[X_i] = O(1)$ since X_i is Poissonian with parameter 2
   4.2 $Cov_[X_i, X_j] = \sqrt{Var[X_i]Var[X_j]}e^{-\Omega(\|p-q\|)} = O(1)e^{-\Omega(\|p-q\|)}$

Other case

If ε is small ($< k^{-1/6}$), then the first method won't work because the value wouldn't exceed the threshold.

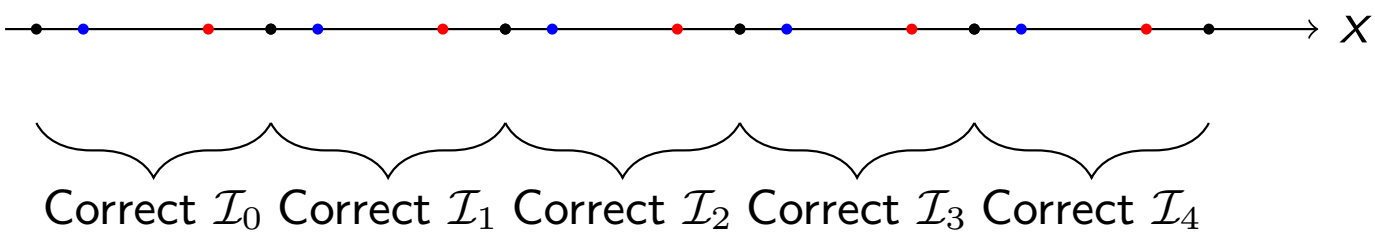
Testing algorithm for small ε

1. Sample $Poi(\frac{k^{4/5}}{\varepsilon^{6/5}})$ samples from $\frac{p+q}{2}$.
2. Let p' and q' be reduced distributions on intervals separated by these samples. They would be nearly uniform then. (W.h.p $\|p', q'\|_2 < O(\frac{1}{\sqrt{n}})$.)
3. Test $p' = q'$ with a tester from part one with sensitivity $\frac{\varepsilon}{C}$ and probability $\frac{1}{C^2}$

Correctness of reduction to tester 2

Claim 1: If $|p - q| \geq \varepsilon$ and at least $\varepsilon/2$ comes from intervals of weight $> 1/m$, then $|p' - q'| \geq \frac{\varepsilon}{C}$.

Proof: We have O(1) samples in each interval. We make a partition from the first and the last sample.



Lower bound (discrete)

$$\text{Why} \frac{t^{4/5}}{\varepsilon^{6/5}}?$$

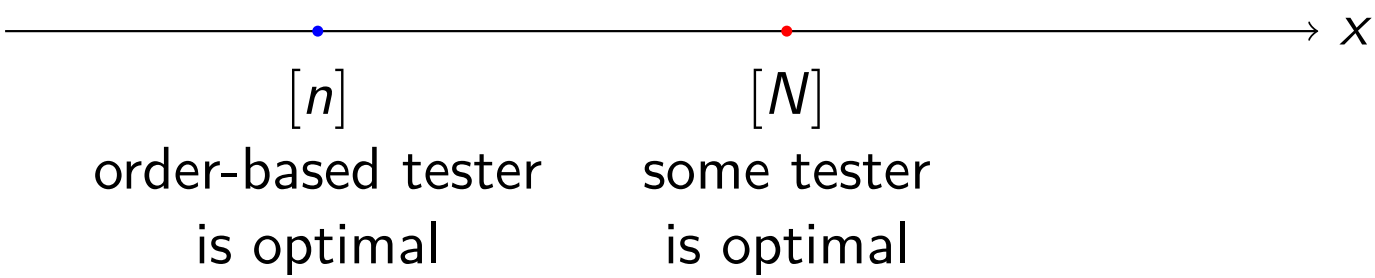
Lower bound

- Optimal tester only depends on the ordering of samples.(Domain blowup.)
- Distributions, which are hard if the tester only looks on the order, exist.

Claim 1:

$\forall n$, $\exists N$ if $\exists A$ working on $[N]$, then $\exists A'$ working on $[n]$, such that it only looks on the ordering on the samples.

In other words: any lower bound on an order-based tester on $[n]$ implies a lower bound on ANY tester on some $[N]$. And a lower bound on $[N]$ implies a lower bound on any $[N* > N]$. (For the same k!)

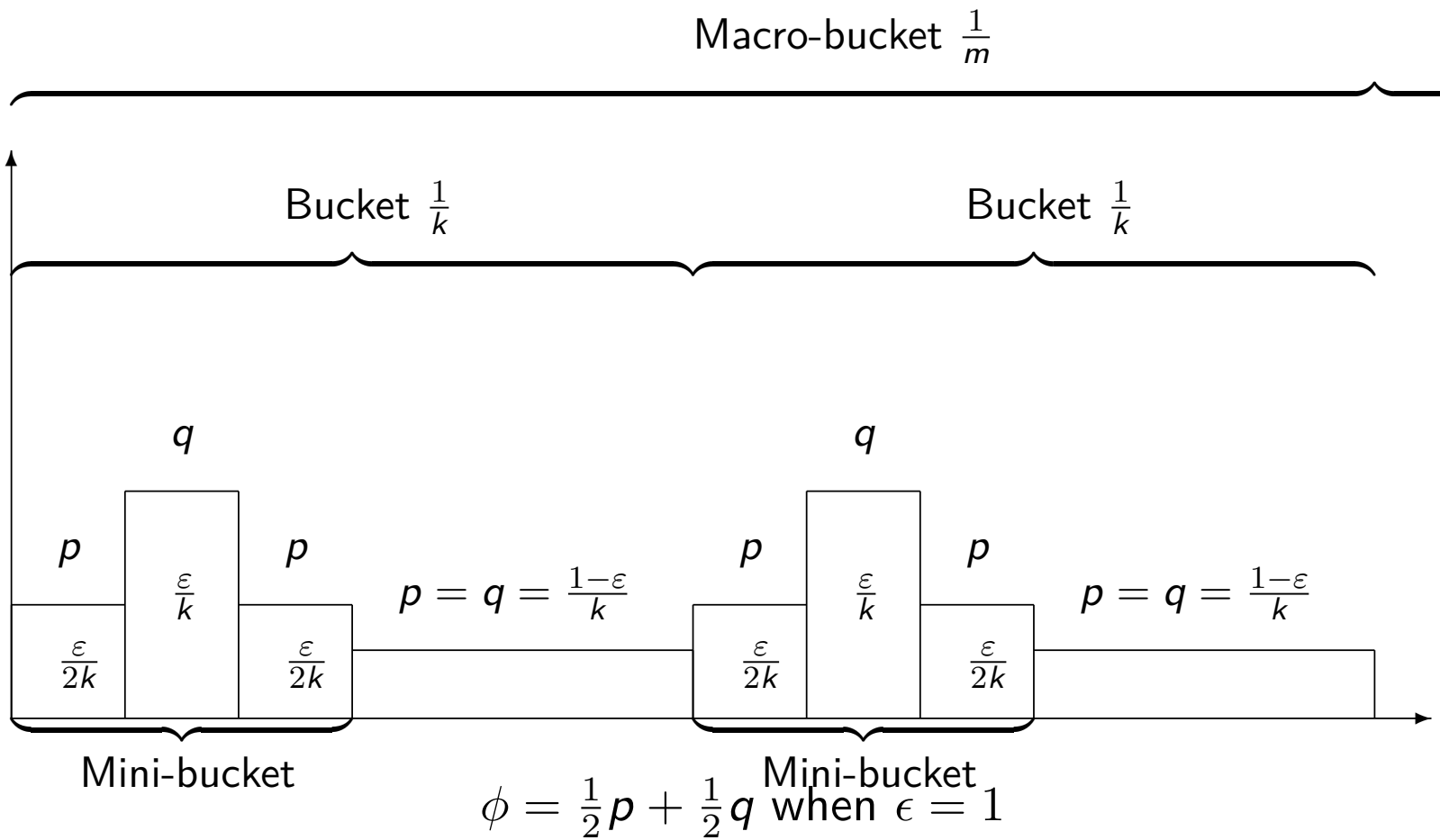


Claim 1: Ramsey Theorem

$\forall m, b, n\ \exists N$ such that $\forall f \binom{[N]}{m} \to b\ \exists S\ (|S| = n) \subset [N]$ so that f has the same value on any $\binom{S}{m}$.

In our case $b = (2^m)^{(2^m)} \leqslant 2^{2^{4m}}$ (all possible functions a tester may use to obtain the result)

Claim 2: distribution



Claim 2: Proof

- Define the random variable X as 1 or 0 with probability $\frac{1}{2}$.
- Define the random variable Y to be ordered sequences of samples from p and q. If $X = 0$, $p = q = \frac{1}{n}$, else p, q as above.

Mutual information between X and Y determines the lower bound on sample size.

Claim 2: Proof

For each macro-bucket, we define Y_i.

$$I(X : Y) \leq \sum_{i=1}^{O(m)} I(X : Y_i)$$

$$I(X, Y_i') \leq O(\frac{m^4 \varepsilon^6}{k^4})$$

Proper algorithm for identity testing

Algorithm Test-Uniformity-$\mathcal{A}_k(q, n, \varepsilon)$
Input: sample access to a distribution q over $[n]$, $k \in \mathbb{Z}_+$ with $2 \le k \le n$, and $\varepsilon > 0$.
Output: "YES" if $q = U_n$; "NO" if $\|q - U_n\|_{\mathcal{A}_k} \ge \varepsilon$.

1. Draw a sample S of size $m = O(\sqrt{k}/\varepsilon^2)$ from q.
2. Fix $j_0 \in \mathbb{Z}_+$ such that $j_0 \stackrel{\text{def}}{=} \lceil \log_2(1/\varepsilon) \rceil + O(1)$. Consider the collection $\{\mathcal{I}^{(j)}\}_{j=0}^{j_0-1}$ of j_0 partitions of $[n]$ into intervals; the partition $\mathcal{I}^{(j)} = (I_i^{(j)})_{i=1}^{\ell_j}$ consists of $\ell_j = k \cdot 2^j$ many intervals with $p(I_i^{(j)}) = 1/(k \cdot 2^j)$, where $p \stackrel{\text{def}}{=} U_n$.
3. For $j = 0, 1, \ldots, j_0 - 1$:
   - 3.1 Consider the reduced distributions $q_r^{\mathcal{I}^{(j)}}$ and $p_r^{\mathcal{I}^{(j)}} \equiv U_{\ell_j}$. Use the sample S to simulate samples to $q_r^{\mathcal{I}^{(j)}}$.
   - 3.2 Run Test-Uniformity-$L_2(q_r^{\mathcal{I}^{(j)}}, \ell_j, \varepsilon_j, \delta_j)$ for $\varepsilon_j = C \cdot \varepsilon \cdot 2^{3j/8}$ for $C > 0$ a sufficiently small constant and $\delta_j = 2^{-j}/6$, i.e., test whether $q_r^{\mathcal{I}^{(j)}} = U_{\ell_j}$ versus $\|q_r^{\mathcal{I}^{(j)}} - U_{\ell_j}\|_2 > \gamma_j \stackrel{\text{def}}{=} \varepsilon_j/\sqrt{\ell_j}$.
4. If all the testers in Step 3(b) output "YES", then output "YES"; otherwise output "NO".

The proper algorithm for equivalence testing

Algorithm Simple-Test-Identity-$\mathcal{A}_k(p, q, \varepsilon)$
Input: sample access to pdf's $p, q : [0, 1] \to R_+$, $k \in Z_+$, and $\varepsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$.

1. Let $m = C \cdot (k^{4/5}/\varepsilon^{6/5})$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size Poi(m) from p and from q respectively.
2. Merge S_p and S_q while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).
3. Compute the statistic Z defined as follows:

$$Z \stackrel{\text{def}}{=} \#\,(\text{pairs of successive samples in } S \text{ coming from the same distribution}) - \#\,(\text{pairs of successive samples in } S \text{ coming from different distributions})$$

4. If $Z > 3 \cdot (\sqrt{m})$ return "NO". Otherwise return "YES".

The proper algorithm for small ε

Algorithm Test-Identity-$\mathcal{A}_k(p, q, \varepsilon)$
Input: sample access to distributions $p, q : [0, 1] \to R_+$, $k \in Z_+$, and $\varepsilon > 0$.
Output: "YES" if $q = p$; "NO" if $\|q - p\|_{\mathcal{A}_k} \geq \varepsilon$.

1. Let $m = Ck^{4/5}/\varepsilon^{6/5}$, for a sufficiently large constant C. Draw two sets of samples S_p, S_q each of size Poi(m) from p and from q respectively.
2. Merge S_p and S_q while remembering from which distribution each sample comes from. Let S be the union of S_p and S_q sorted in increasing order (breaking ties randomly).
3. Compute the statistic Z defined as follows:

$$Z \stackrel{\text{def}}{=} \#\left(\begin{array}{c}\text{pairs of successive samples in } S \\ \text{coming from the same distribution}\end{array}\right) - \#\left(\begin{array}{c}\text{pairs of successive samples in } S \\ \text{coming from different distributions}\end{array}\right)$$

4. If $Z > 5\sqrt{m}$ return "NO".
5. Repeat the following steps $O(C)$ times:
   (a) Draw Poi(m) samples from $(p + q)/2$.
   (b) Split the domain into intervals with the interval endpoints given by the above samples. Let p' and q' be the reduced distributions with respect to these intervals.
   (c) Run the tester of the first part on p' and q' with error probability $1/C^2$ to determine if $\|p' - q'\|_{\mathcal{A}_{2k+1}} > \varepsilon/C$. If the output of this tester is "NO", output "NO".
6. Output "YES".