Testing Identity of Structured Distributions

Ilias Diakonikolas [1] Daniel Kane [2] Vladimir Nikishkin [1]

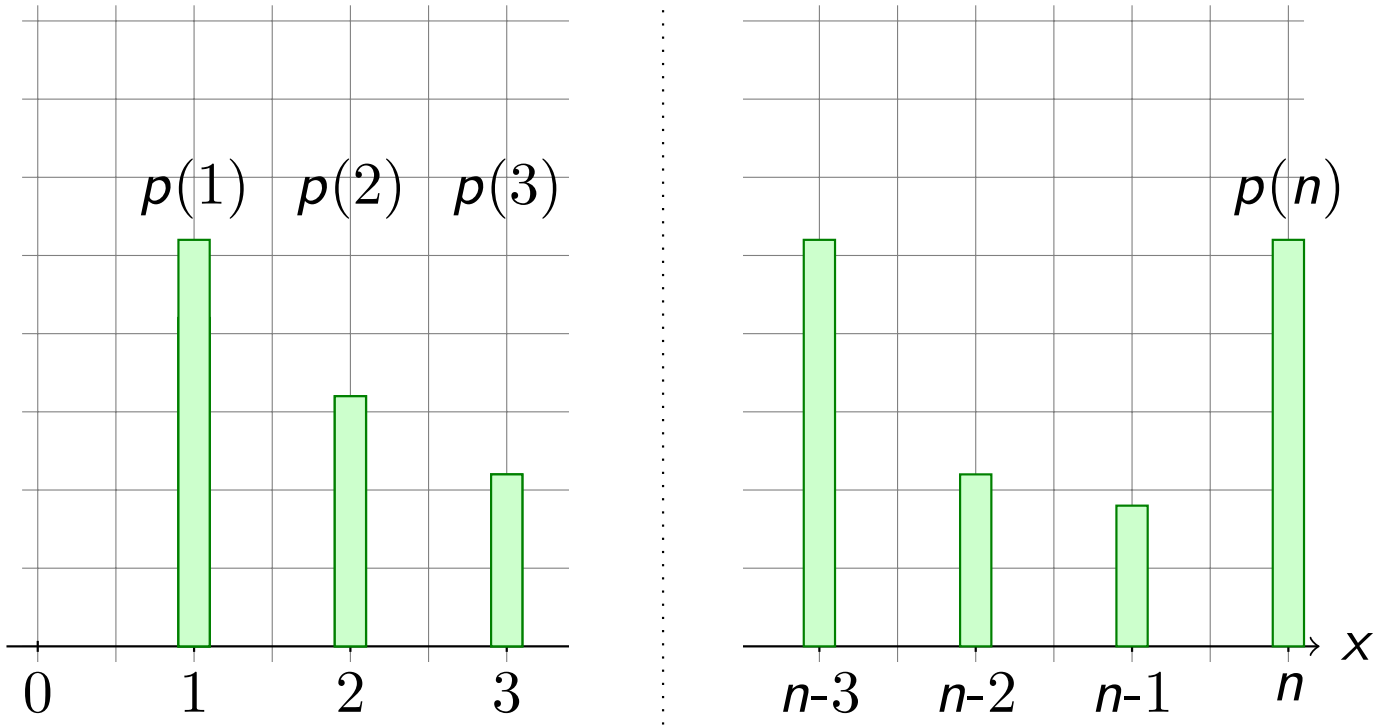
[1]University of Edinburgh

[2]University of California, San Diego

January 2015, SODA 2015, San Diego

What this talk is about

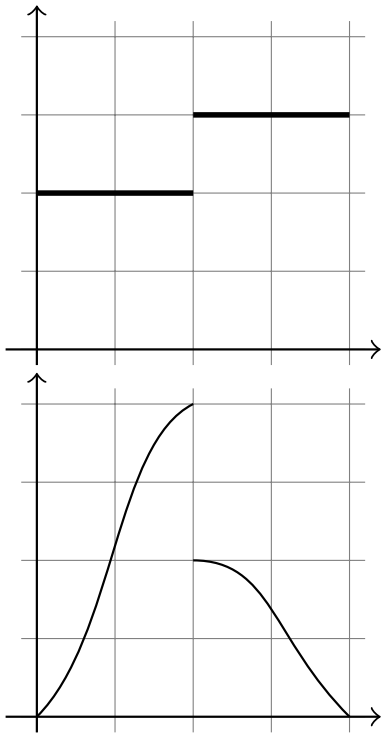
Discrete probability distributions over $[n] = \{1, 2, ..., n\}$.



Structured distributions

Shape constrained distributions:

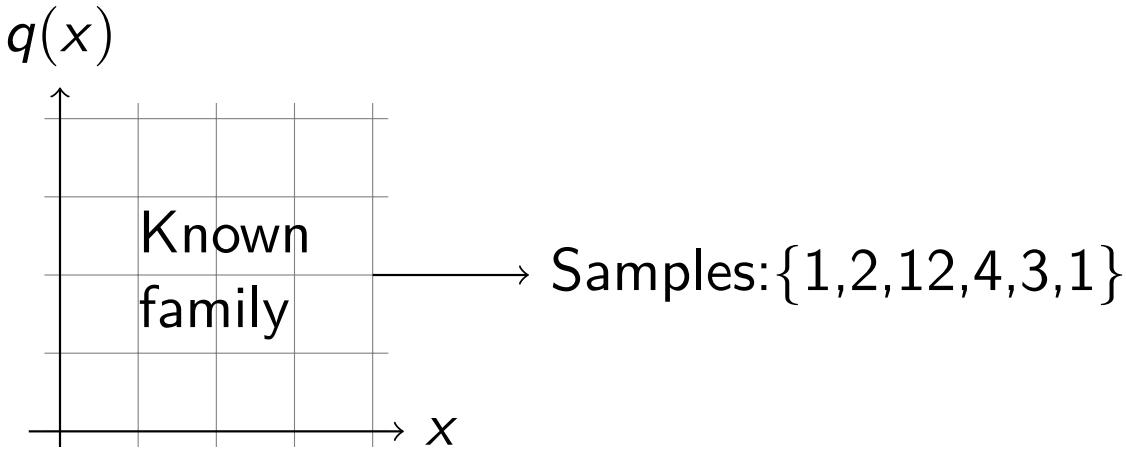
- Piecewise-constant.
- Piecewise-polynomial.
- Log-concave.
- Monotone Hazard Rate.
- Mixtures of the above.



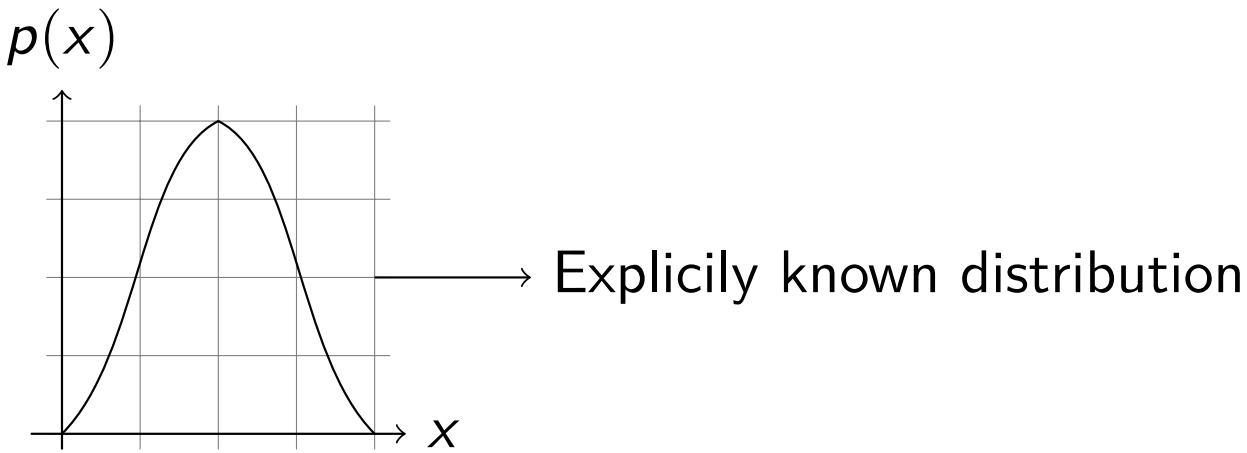
Problem definition: identity testing

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

- q :



- p :



Prior research

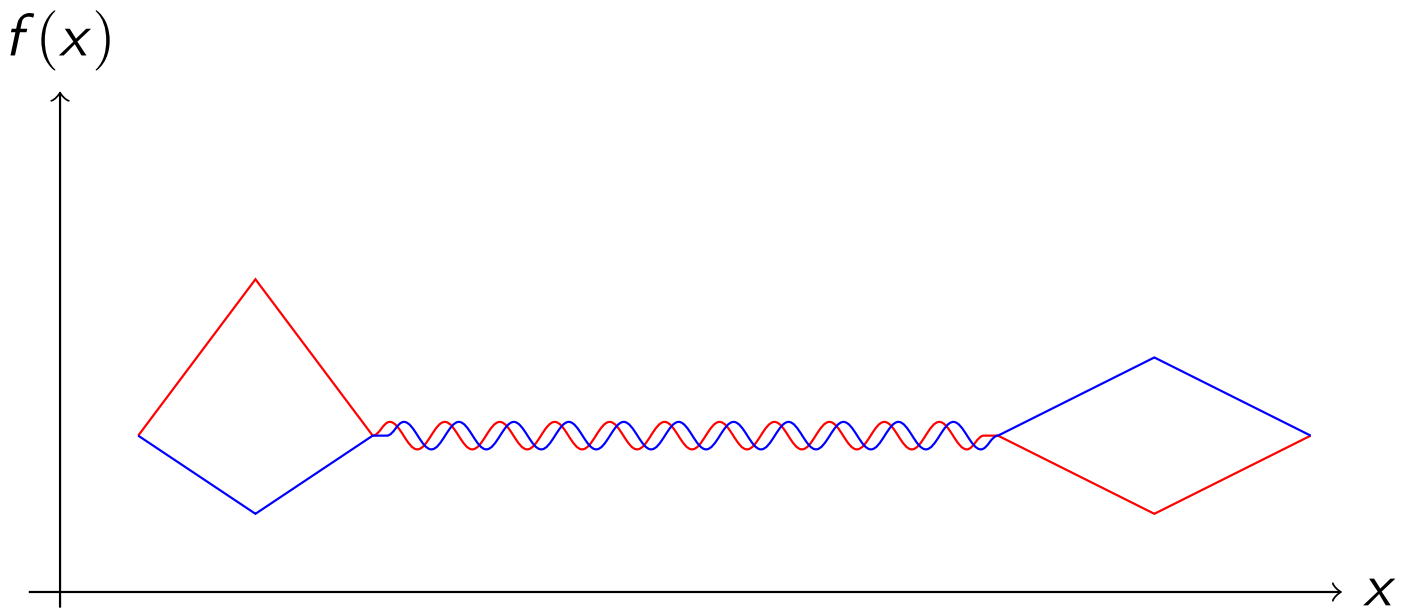
- Statistics
  1. Hypothesis testing: [Pea1900], [NP1933]
  2. Shape constrained inference: [Gre56], [Bru58], [Rao69]
- TCS
  1. Property testing: [RS96], [GGR98]
  2. Distribution testing: [GR00], [Batu2000], [CDVV14], [Pan08], [Val11], [VV11], [ADJ+11]
  3. Learning and testing structured distributions: [DDS12] (k-modal learning), [DDSVV13] (k-modal testing), [CDSS14] (piecewise polynomial learning)

Our Results

- Unified approach for L_1 identity testing of structured distributions.
- Sample optimal testers for broad classes of structured distributions via a single algorithm.

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

Intuition



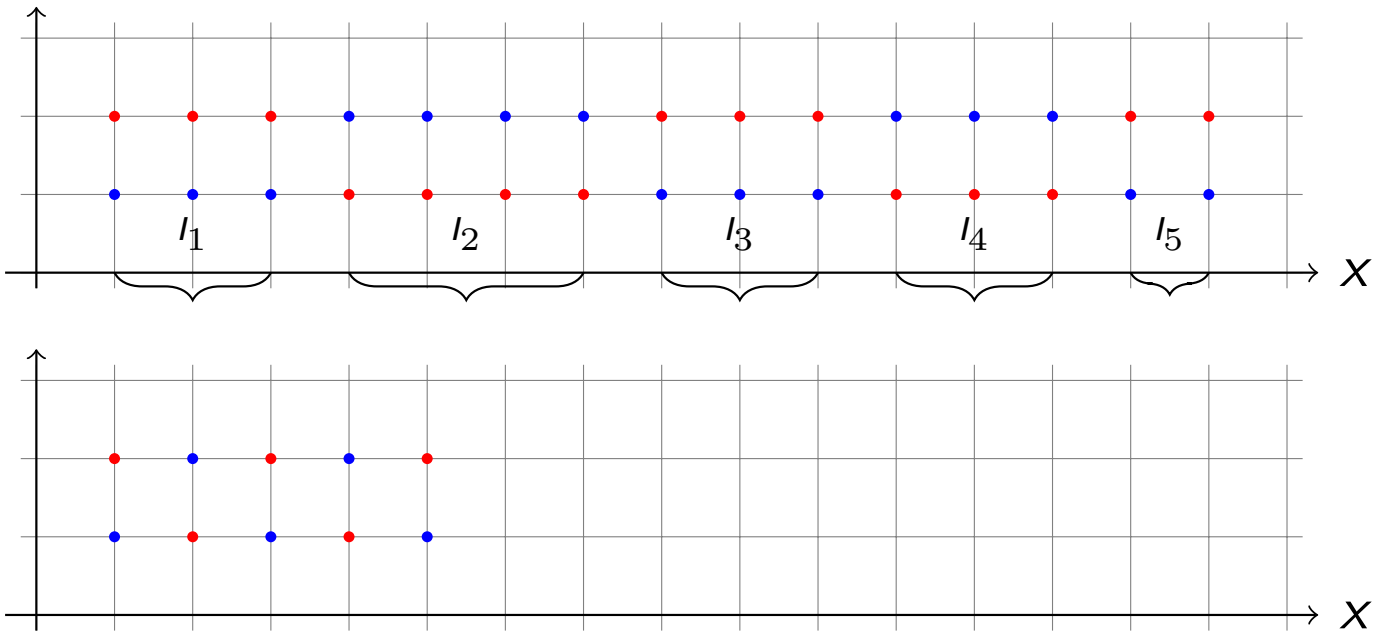
4 essential crossings.

Definition of A_k distance

A_k distance - structured equivalent of L_1 .

Let $\mathcal{J}_k$ be the set of all partitions of the domain into k intervals.

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



Main result

Optimal algorithm for testing identity under the A_k distance.

Theorem

(Main) There is an algorithm which for all discrete distributions p, q tests whether $p = q$ vs $\|p - q\|_{A_k} \geqslant \varepsilon$ using $O(\sqrt{k}/\varepsilon^2)$ samples.

Sample complexity is information theoretically optimal up to a constant factor.

Corollary of Main Result

Question: How many samples do we need to L_1 test a distribution family $\mathcal{C}$ with this algorithm?

Answer: Find **minimal** k such that for any $p, q \in \mathcal{C}$ it holds:

$$L_1(p, q) \leqslant A_k(p, q) + \frac{\varepsilon}{2}$$

Then we can test L_1 distance with our protocol using $O(\sqrt{k}/\varepsilon^2)$ samples.

Corollary

Let $\mathcal{C}$ be a distribution family that is $\varepsilon/2$-approximated in L_1 norm by t-piecewise degree-d polynomials. Then there is an L_1 identity tester for $\mathcal{C}$ using $O(\sqrt{t(d+1)}/\varepsilon^2)$ samples.

Applications of Main Result

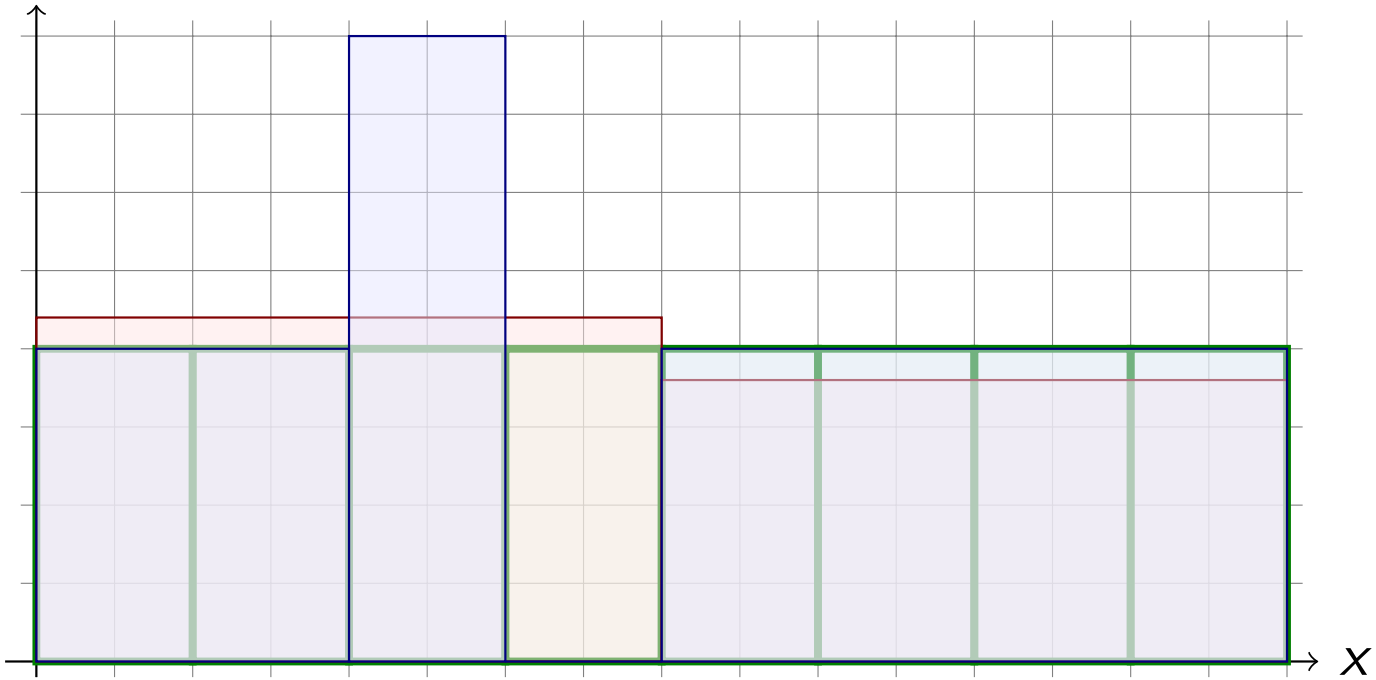
Class of Distributions over $[n]$	**Our upper bound**	**Previous work**	***k***
t-piecewise constant	$O(\sqrt{t}/\varepsilon^2)$	$O(t/\varepsilon^2)$ [CDSS14] (learn)	$O(t)$
t-piecewise degree-d polynomial	$O\left(\sqrt{t(d+1)}/\varepsilon^2\right)$	$O\left(t(d+1)/\varepsilon^2\right)$ [CDSS14] (learn)	$O(t(d+1))$
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)	$O(\ell \log(n/\varepsilon) \cdot 1/\varepsilon)$

Testing Algorithm Overview

1. Reduction of A_k identity testing to A_k uniformity testing
   - Stretch the domain
   - Simulate samples over stretched domain

2. Algorithm for A_k uniformity testing
   - Apply L_2 uniformity tester to various equipartitions
   - Use a strong property of uniformity testing under L_2 norm

L_2 to L_1 testing (Tall vs. wide)



Testing uniformity in L_2 norm

Theorem

Let q be a discrete distribution over $[m]$. There is an algorithm that draws $O(\frac{\sqrt{m}}{\varepsilon^2})$ samples from q and tests whether $q = U_m$ versus $L_2(q, U_m) \geq \varepsilon/\sqrt{m}$.

[GR00] achieved above guarantee with $O(\frac{\sqrt{m}}{\varepsilon^4})$ samples.

Our A_k uniformity testing algorithm applies above algorithm as a black box for various equipartitions of the domain.

Uniformity testing algorithm under A_k distance

Initialise the counter $i = 0$, sensitivity parameter $\delta = \varepsilon$.

1. Partition the domain of q' into i equal intervals and make a reduced distribution q_i.
2. $\texttt{TEST}(q_i, \delta)$
3. Increase i and δ exponentially.
4. `GOTO` 2 or terminate if loop limit reached.

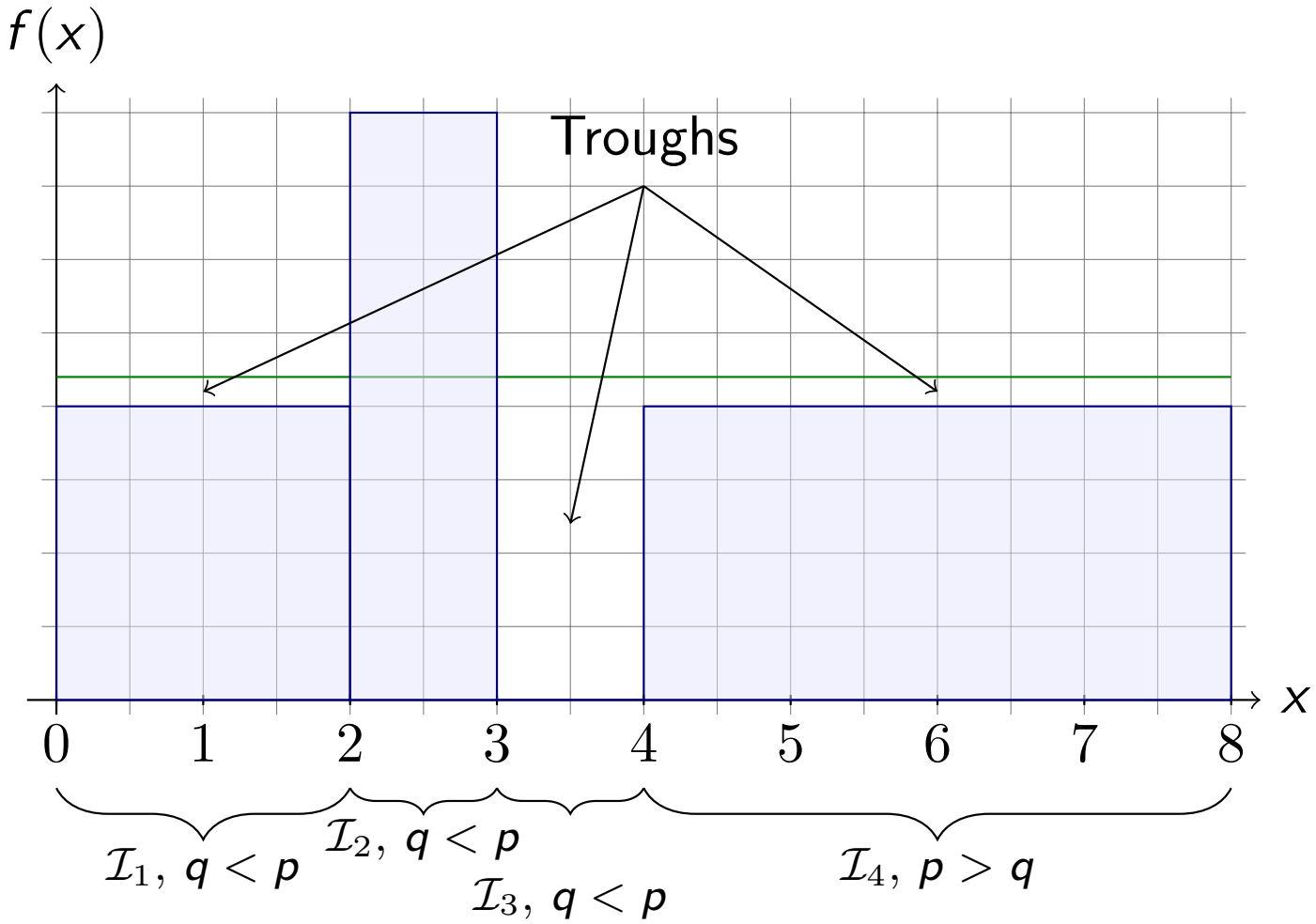
If all tests succeed, consider p and q equal.

Sketch of the Proof

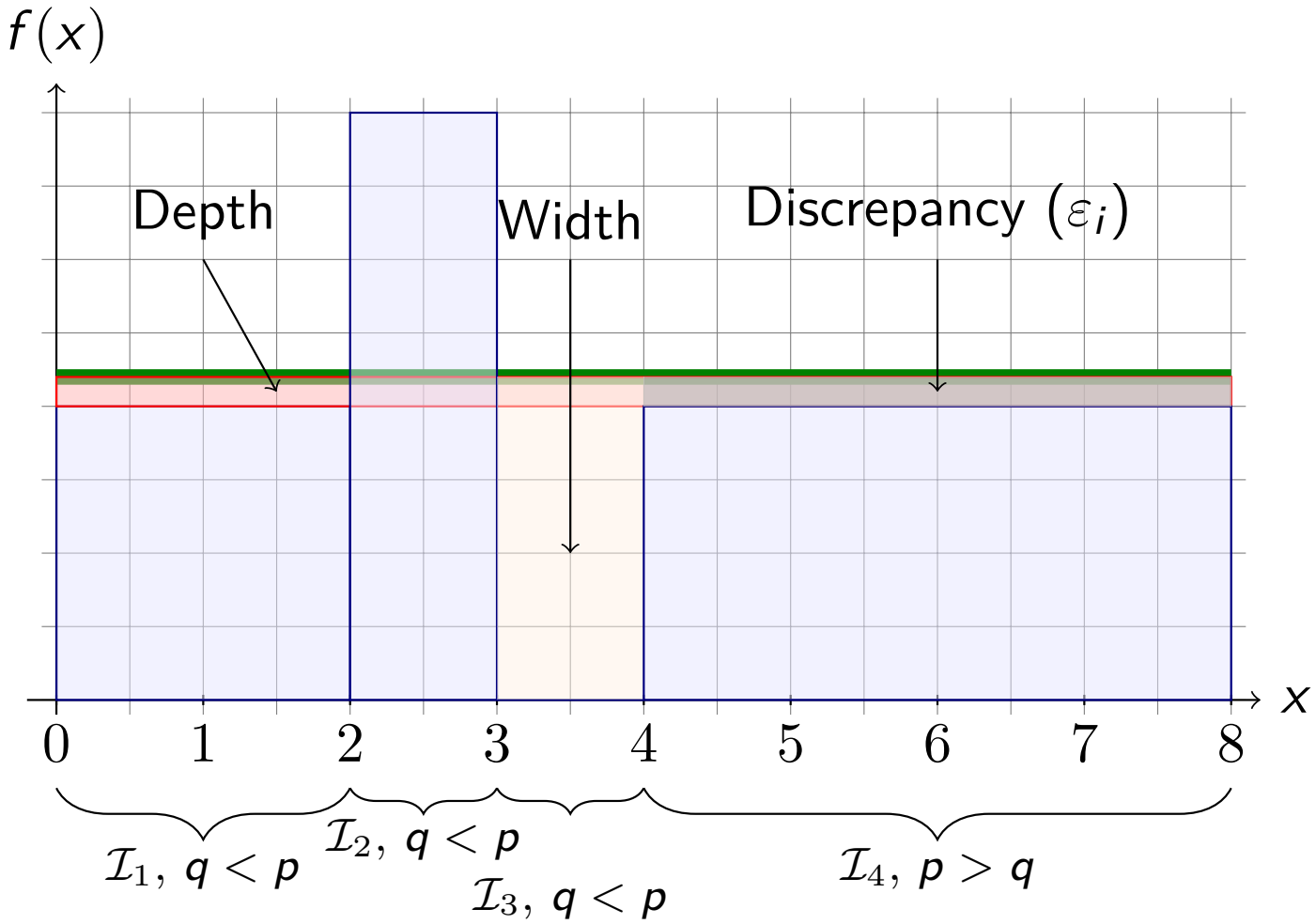
We show that if U and q indeed are far, then in at least one of the partitions the distance between U and q_i is exceeds the allowed threshold.

That is, the sum of discrepancies over all tests is bigger than the sum of all discrepancies possibly missed by all testers. Thus at least one tester will "catch" the discrepancy.

Proof: Illustration



Though structure



Observations

- For every though I_*, width $\leqslant 1/k$.
- Discr$(I_*) \leqslant$ width(I_*)
- We can ignore small thoughs and assume $\varepsilon/(20k) \leqslant$ width$(I_*) \leqslant 1/k$.
- There is a partitioning level j_i such that at least one of its intervals J is contained in I_* and is at least a quarter length.

Therefore we know that for at least one level of reductions, every trough contributes at least $\frac{\varepsilon}{320k}$Discr to the L_2 distance.

The lower bound for the sum of the discrepancies over all levels is greater or equal to than $\frac{\varepsilon^2}{800k}$

Whereas the sum of maximal errors missed by the testers is: $\frac{\varepsilon^2}{6400k} \sum_{j=0}^{j_0-1} 2^{-j/4} < \frac{\varepsilon^2}{800k}$.

Therefore one of the testers will find the discrepancy with high probability.

Conclusion/Open problems

1. This work: Identity to a known distribution
2. Open problems
   - Identity between two unknown structured distributions
   - L_1 distance estimation

Thank you!