Frequency Distribution testing

Michel de Rougemont

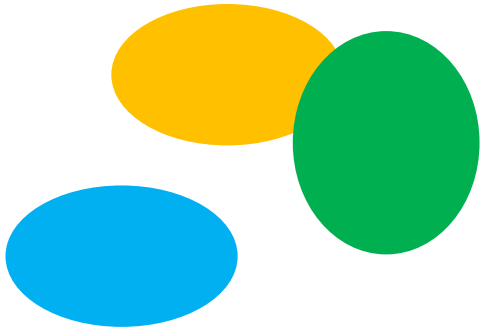
University Paris II & IRIF CNRS

(Joint work with Claire Mathieu)

Motivations

1. Decide very quickly Statistical constraints

2. SAT panorama (n variables with densities, m clauses)



Heuristics adapted for each area, associated with a statistical constraint.

3. Non worst-case complexity

4. L.L.M. How to represent $P(t_k \mid t_1 , t_2 ,\ldots\ldots t_{k-1})$

How is the distribution given?

1. Distribution as an Oracle
   - Sample D : output (a,b) with probability p(a,b)
   - Sample a with probability p(a | b)

2. Streaming data
   - Stream of objects: a,b,a,c,a,a,b……..
   - Stream of texts: all the text on the web…….

Distribution D defined by the frequencies of letters, words,…. Reservoir sampling gives samples on D.

Standard problems

Domain of size n

1. Is the distribution D close to:
   - A fixed distribution (power law, for example)
   - Uniform distribution

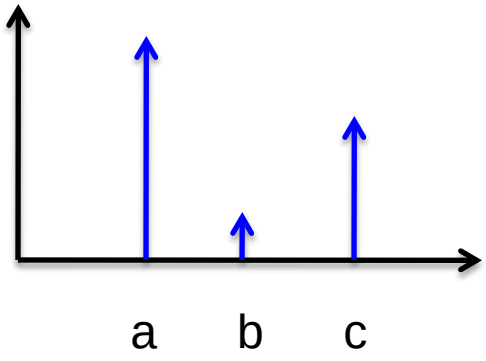
2. Is $D = D_1 * D_2$?

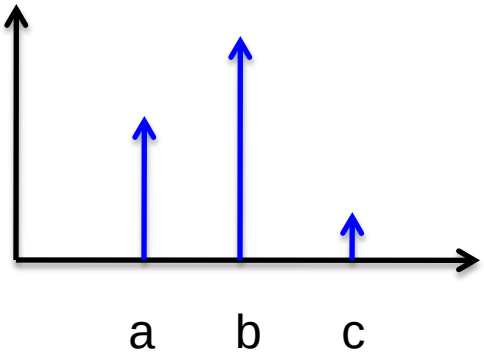
   $D_1 = D_2$?

Frequency distributions

Stream of objects: a,c,a,c,a,a,b,c ……..

c,a,b,b,a,b,b,a……..

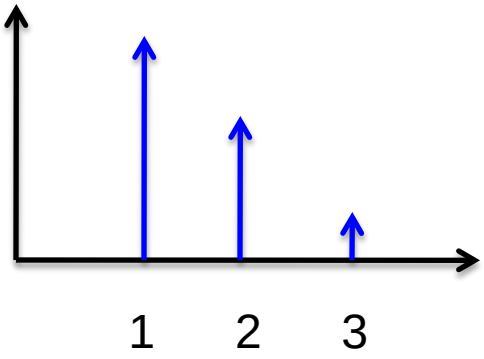




Frequency distribution:

Domain={1,2,…n}

Decreasing function



Plan

1. Distribution Testing
   - Total variation distance
   - Is D close to the Uniform distribution

2. Frequency Distribution Testing
   - Relative Frechet distance
   - Is D close to the Uniform distribution: $\Omega(n)$ space
   - Smooth Decreasing distributions: $O(\log^2 n)$ space

Property Testing framework: algorithm A s.t.

- If D=Uniform distribution, A(D)=1
- If D is ε-far from the Uniform distribution, Prob [A(D)=0]> 1-δ

1. Uniformity testing (Cannone 2023)

	Sample complexity	Notes	References
Collision-based	$\frac{k^{1/2}}{\varepsilon^2}$	"Natural"	Goldreich and Ron, 2000; Diakonikolas *et al.*, 2019b
Unique elements	$\frac{k^{1/2}}{\varepsilon^2}$	Low sensitivity $\varepsilon \gg 1/k^{1/4}$	Paninski, 2008
Modified χ^2	$\frac{k^{1/2}}{\varepsilon^2}$	(None)	Valiant and Valiant, 2017; Acharya *et al.*, 2015; Diakonikolas *et al.*, 2015
Empirical distance to uniform	$\frac{k^{1/2}}{\varepsilon^2}$	Low sensitivity	Diakonikolas *et al.*, 2018
Random binary hashing	$\frac{k}{\varepsilon^2}$	Suboptimal, but fast	Acharya *et al.*, 2020d
Bipartite collisions	$\frac{k^{1/2}}{\varepsilon^2}$	Tradeoff possible	Diakonikolas *et al.*, 2019a
Empirical subset weighting	$\frac{k^{1/2}}{\varepsilon^2}$	Tradeoff possible $\varepsilon \gg 1/k^{1/4}$	Acharya *et al.*, 2022

|D|=k

Collision-based tester

Sample $X_1, X_2, \ldots X_m$
Count Collisions

Compute: $Z = \frac{\sum_{1 \le s \le t \le n} 1\{X_s = X_t\}}{n(n-1)/2}$ $\quad E(Z) = \|p\|_2^{\,2}$

General: $\|p-q\|_2 \le \frac{1}{2}\|p-q\|_1 \le \frac{\sqrt{n}}{2}\|p-q\|_2$

If q is uniform: $\|p-q\|_2^{\,2} = \sum_i (p(i) - 1/n)^2 = \|p\|_2^{\,2} - 1/n$

$\|p-q\|_1 \ge \varepsilon \quad \rightarrow \quad \|p\|_2^{\,2} \ge (1 + 4\varepsilon^2)/n$

Study Var(Z), apply Chebyshev inequality

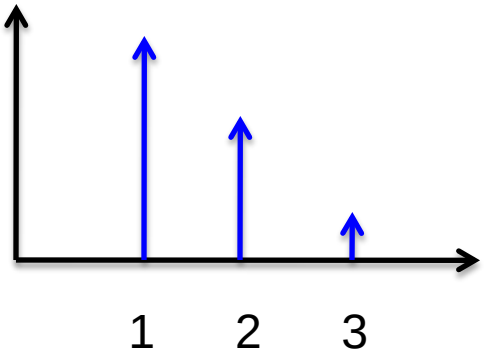
2. Frequency distributions

Stream of objects: a,c,a,c,a,a,b,c ……..

(a,b),(a,c),(b,c)…….

Streaming algorithm in space **O(poly(log n))** to decide if the frequency distribution g of the stream is close to a given distribution f ?

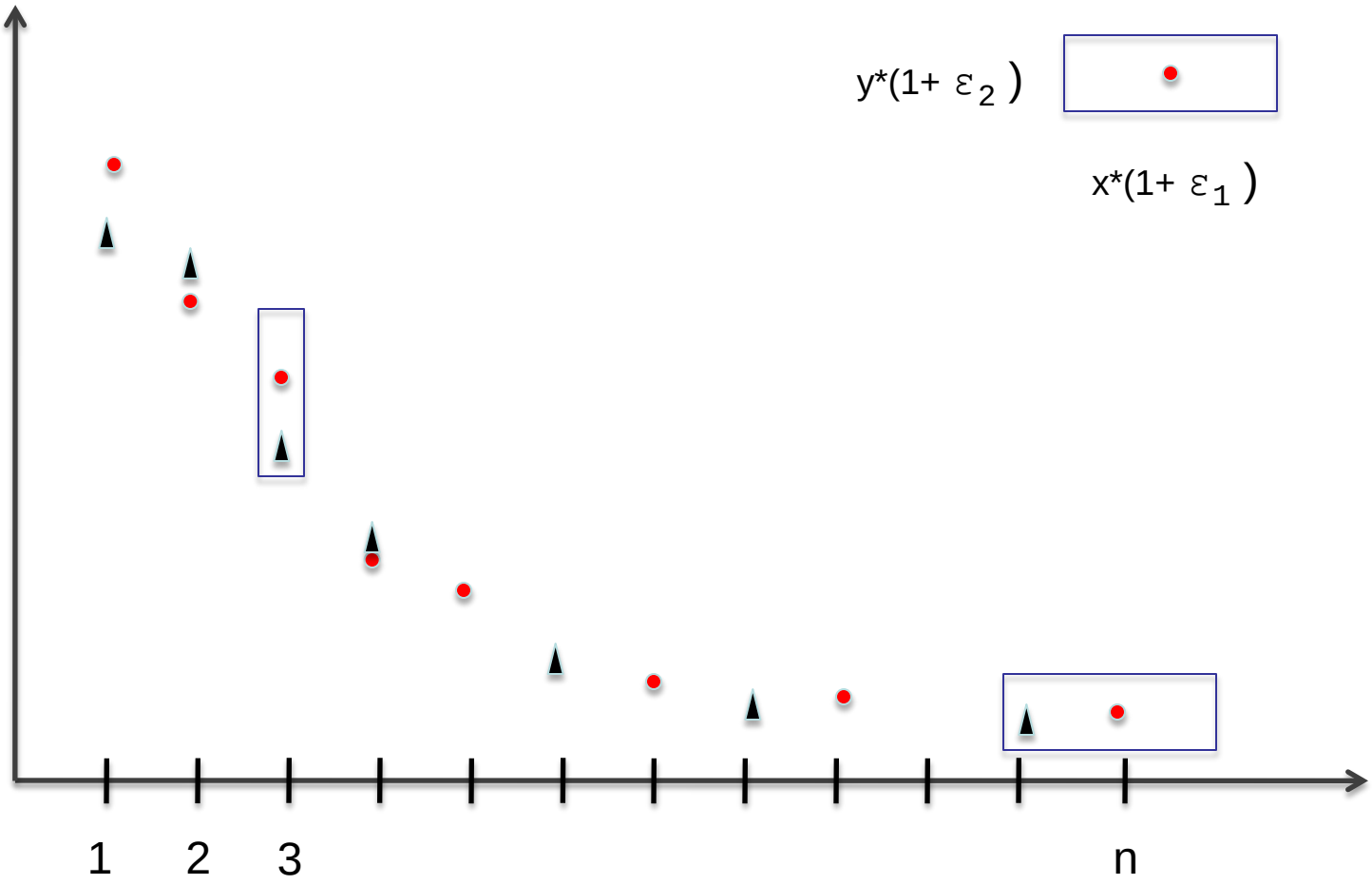
Frequency distribution (absolute values):



Distances between distributions

1. L_1 , L_2 … L_p
2. EMD
   - Earth-moving distance
3. Fréchet
   - Absolute/relative
4. Kullback-Liebler Divergence

Fréchet relative distance



Degree distributions: f(i)=#v: degree(v)=i

Are ▲ and ● close ?

Results (n items)

1. Is g close to the uniform distribution f ?

   $\Omega(n)$ space

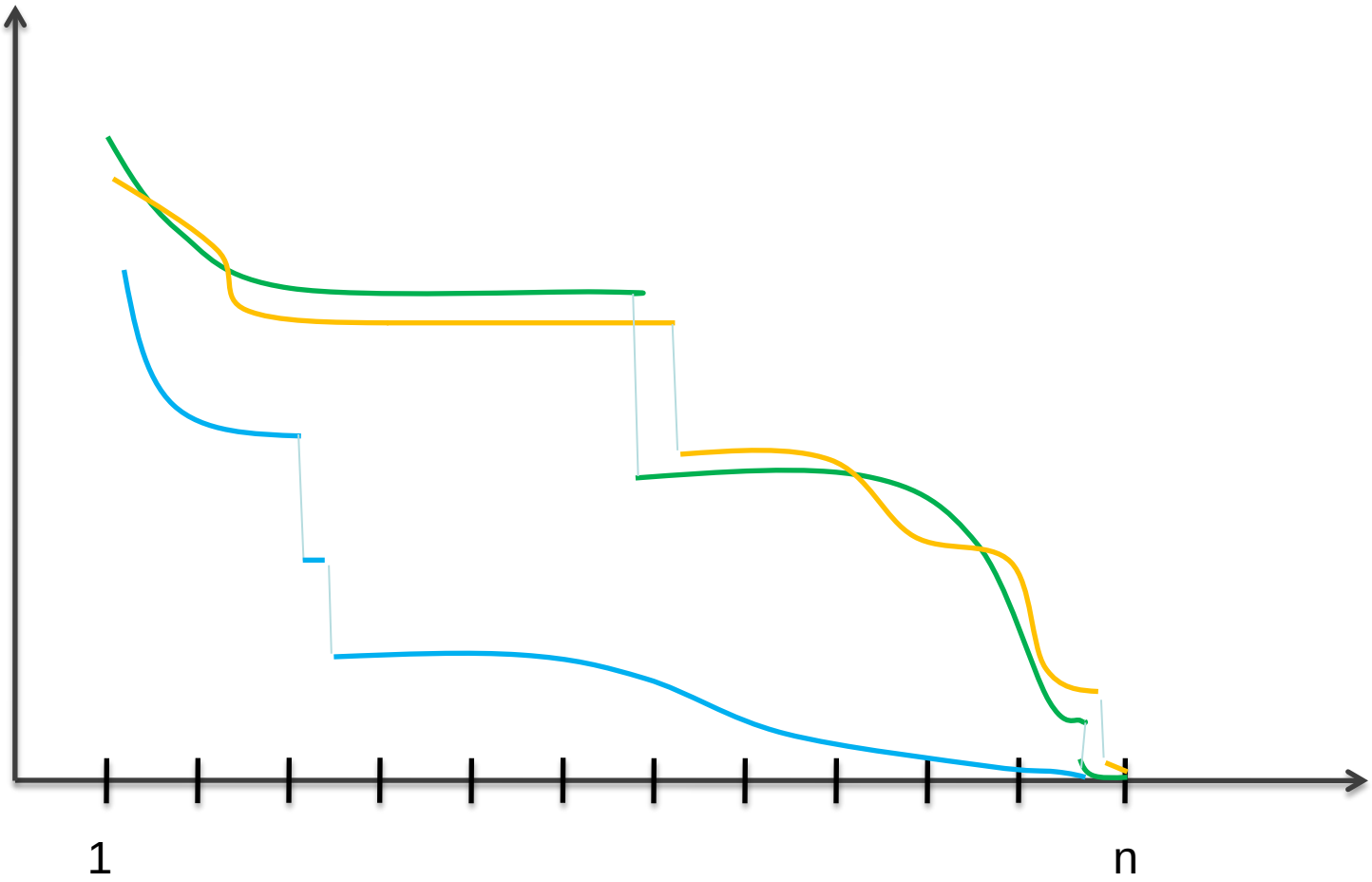
   Communication Complexity (reduction from Index)

2. Smooth and decreasing distributions f

   - Step compatible (Smooth): every point belongs to a complete Frechet rectangle
   - $\gamma - decreasing$: $f(\gamma.t) \leq \frac{f(t)}{2}$

Main result: Streaming tester with $O(\log^2 n)$ space

Smooth distributions

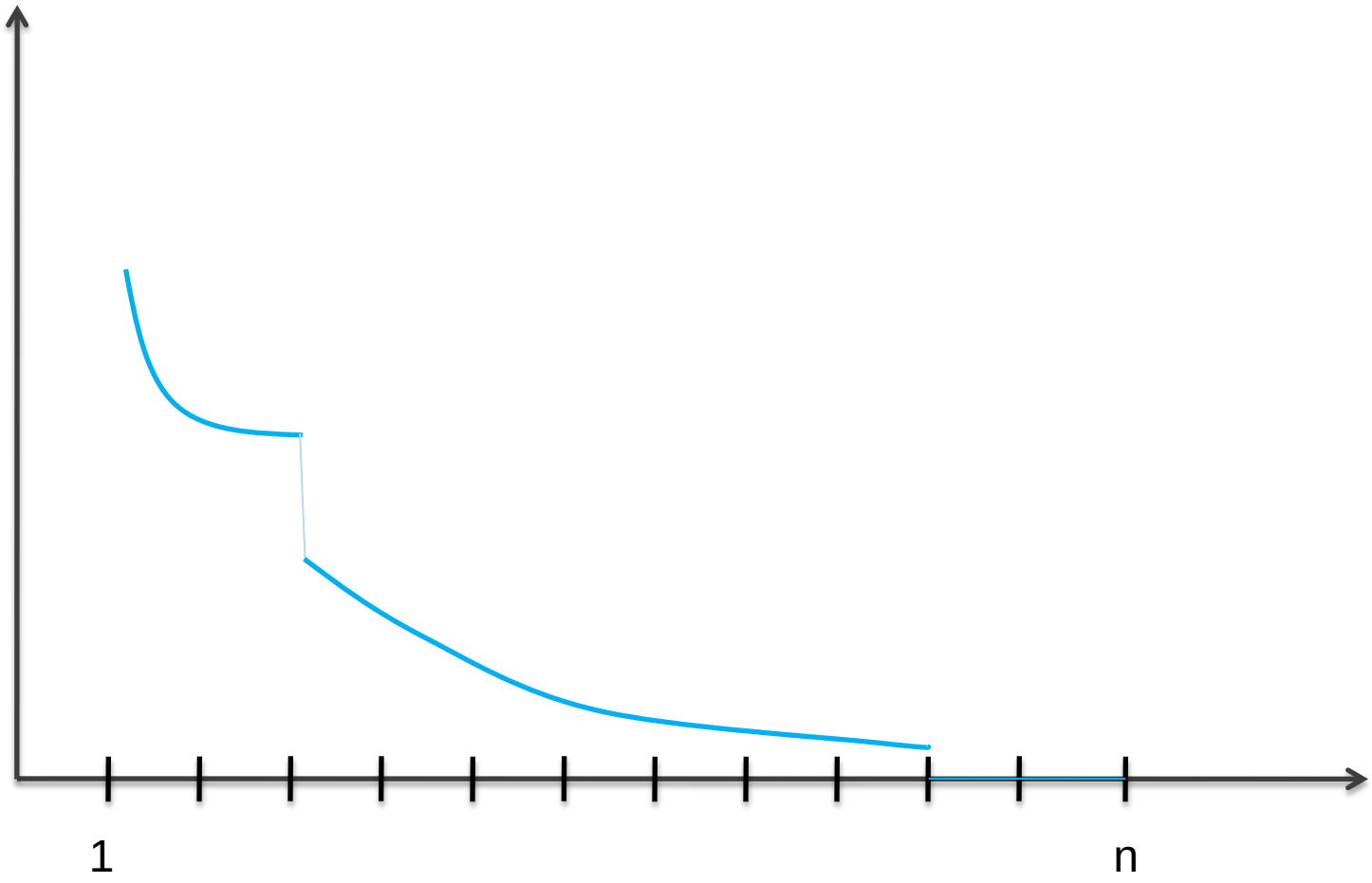


Smooth

Non smooth

Close

Smooth and γ-decreasing



Frequent items

1. Deterministic algorithms
   - Misra-Gries
   - Spacesaving (K=3, additive error)

aaabbc

Counter	Element	ε
3	a	0
2	b	0
1	c	0

d

Counter	Element	ε
3	a	0
2	b	0
2	d	1

$a \in K$: $counter(a) - \varepsilon(a) \leq F(a) \leq counter(a)$

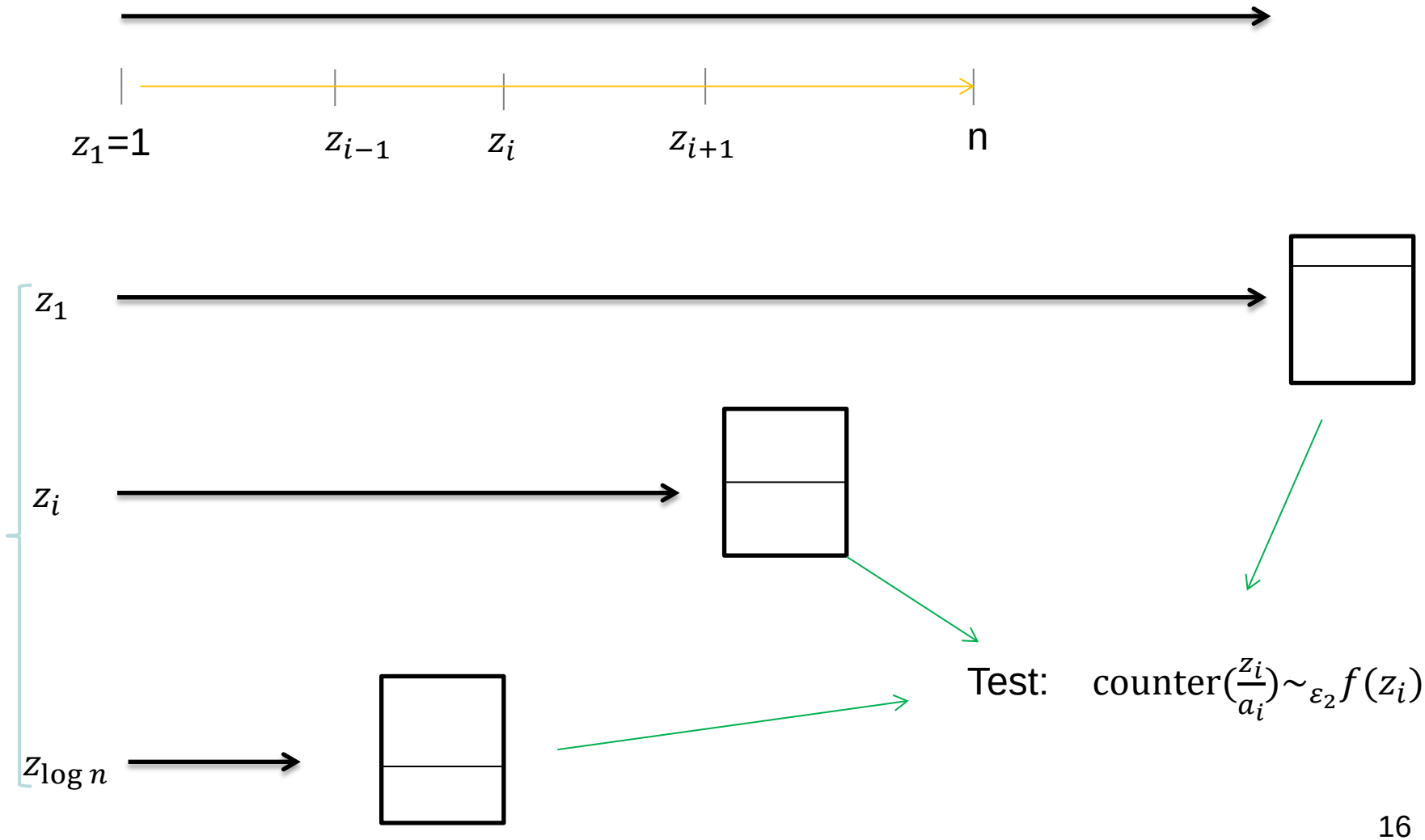
$e\ not\ in\ K$ $F(e) \leq Min\ counter$

2. Randomized algorithms
   - Count-Min-Sketch

Simplified Tester: log n substreams with Spacesaving

$z_i = \left\lfloor \left(1+\varepsilon_1{}^2\right)^i \right\rfloor$ $\qquad$ $a_i = \left\lceil \frac{z_i.\varepsilon_1{}^2}{\log\log n} \right\rceil$

Random hash function h_i : D→[1…a_i] $\qquad$ Keep elements e s.t. h_i (e)=1

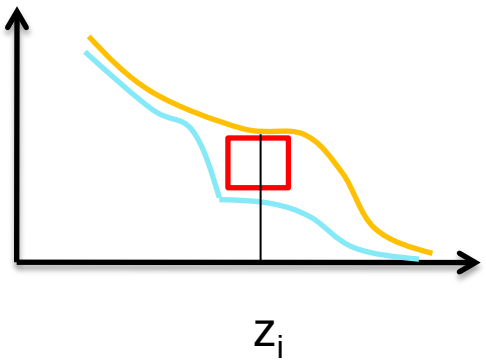


Tester

1. Details
   - Ignore z_i close to the discontinuities of f
   - Spacesaving table of size O(log n . log log n)
   - Exact counting if $\varepsilon_2 . f(z_i) \leq f(n)$

2. Key points
   - If f is γ -decreasing, Spacesaving guarantees relative errors.
   - If f,g are smooth and $\neg\left(f \sim_{3\varepsilon_1, 3\varepsilon_2} g\right)$

     $\rightarrow \quad \exists\ (\varepsilon_1, \varepsilon_2)$ separating rectangle
     - Some z_i will witness the separating rectangle
     - The tester will reject with h.p.



Extensions

1. Tuples in dimension d: $(a_1, a_2, \ldots. a_d)$

   – Domain size: n^d
   – Frequency of the Marginals: hash on the projections

2. Questions
   – Independence of subsets
   – Marginals are close

$$D = D_1 * D_2 \ ?$$
$$D_1 = D_2$$

Conclusion

1. Relative Frechet distance

2. Frequency distribution testing
   - Uniform testing is hard
   - Smooth and decreasing frequency distributions are « easy »

     (Zipf, Power law, …..)

3. Extensions to other properties