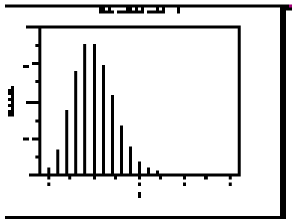
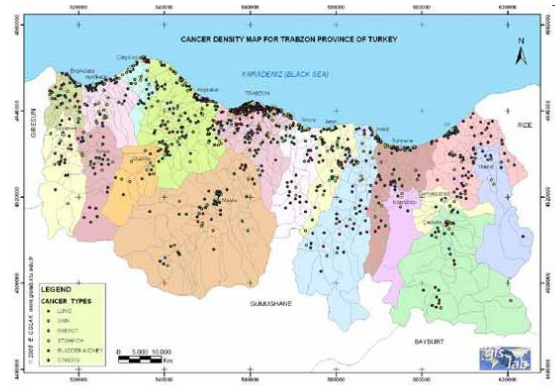
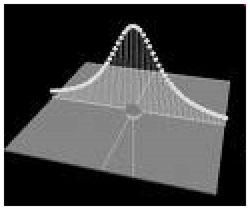
Testing properties of distributions

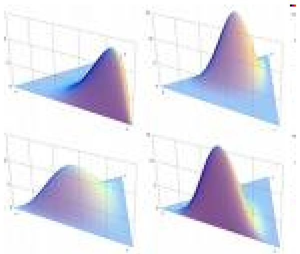
Ronitt Rubinfeld

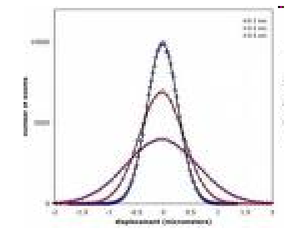
Tel Aviv University and MIT



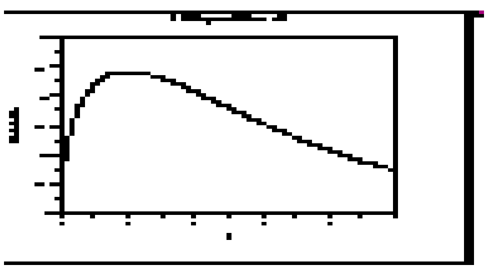


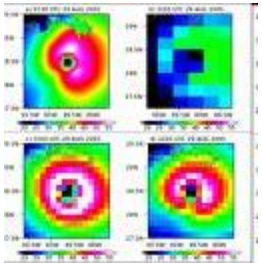


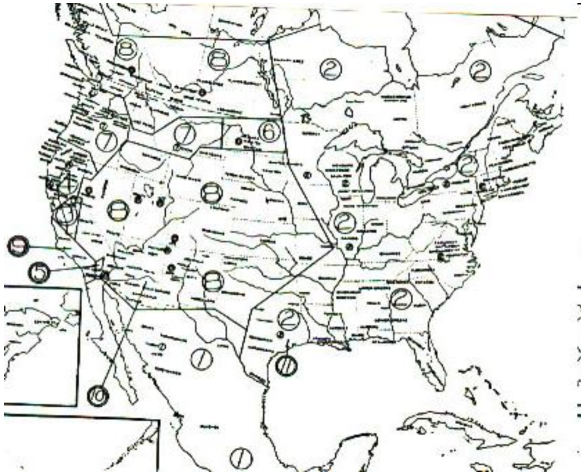




Distributions are everywhere

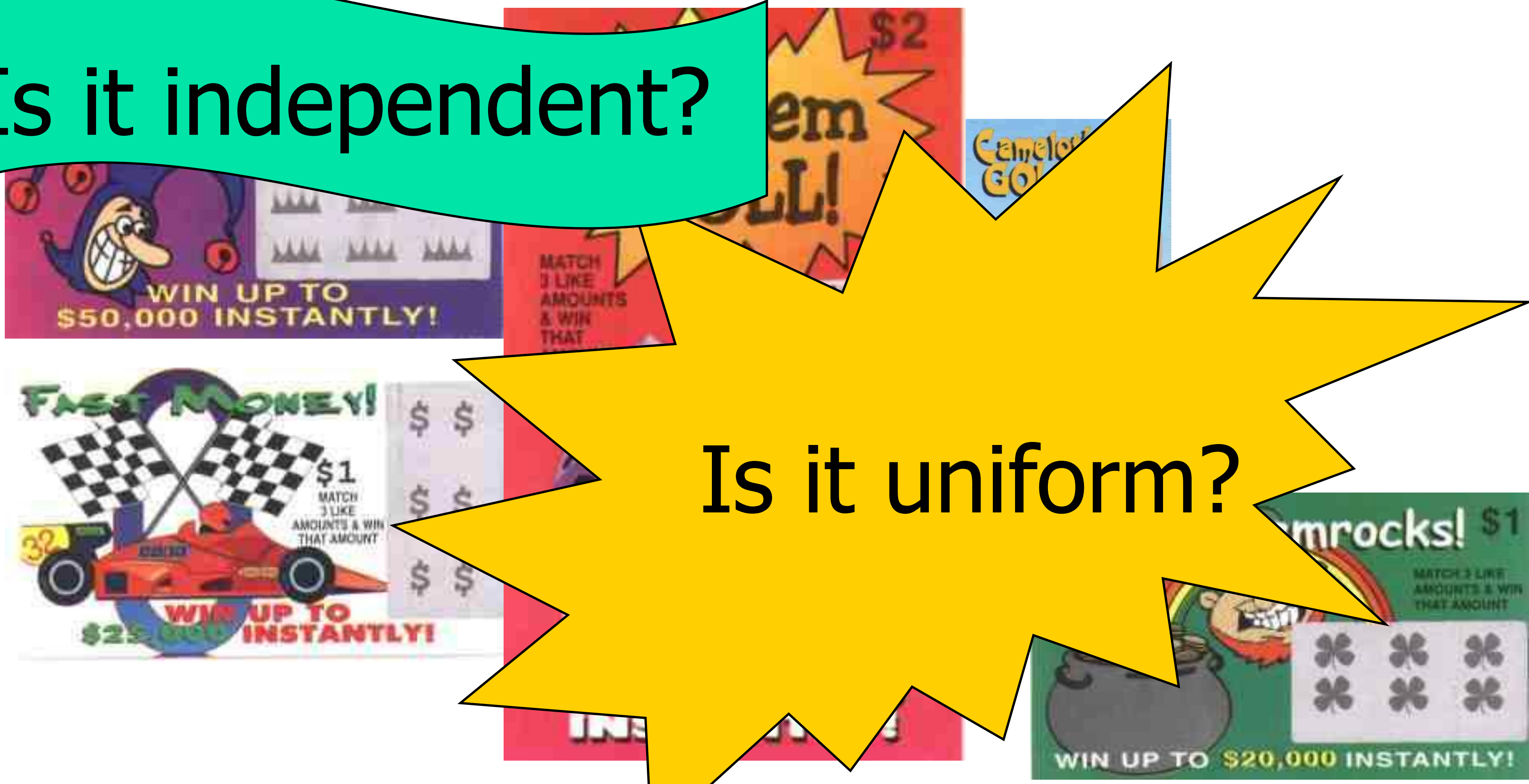






What properties do your distributions have?

Play the lottery?



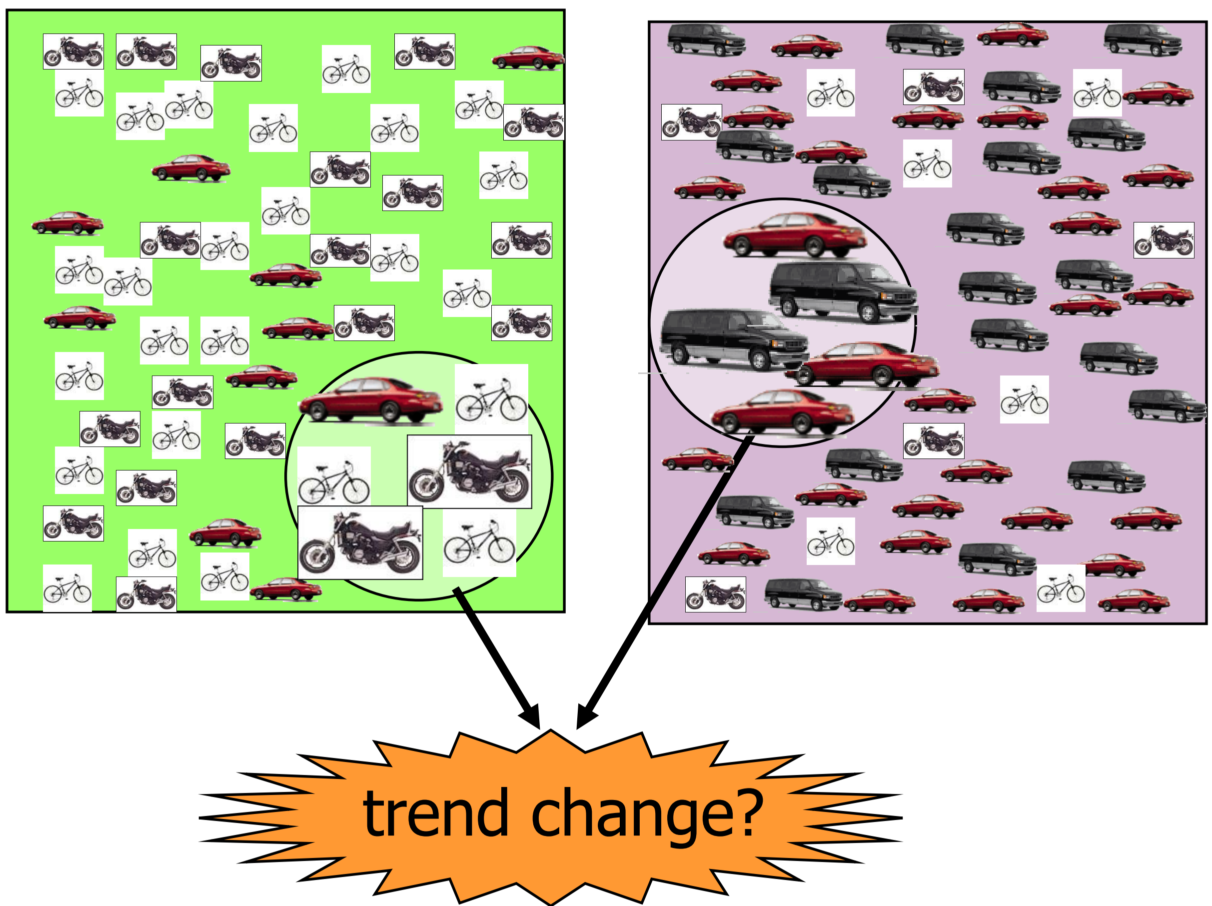
Is it independent?

Is it uniform?

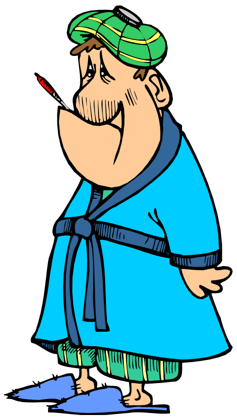
Testing closeness of two distributions:

Transactions of 20-30 yr olds

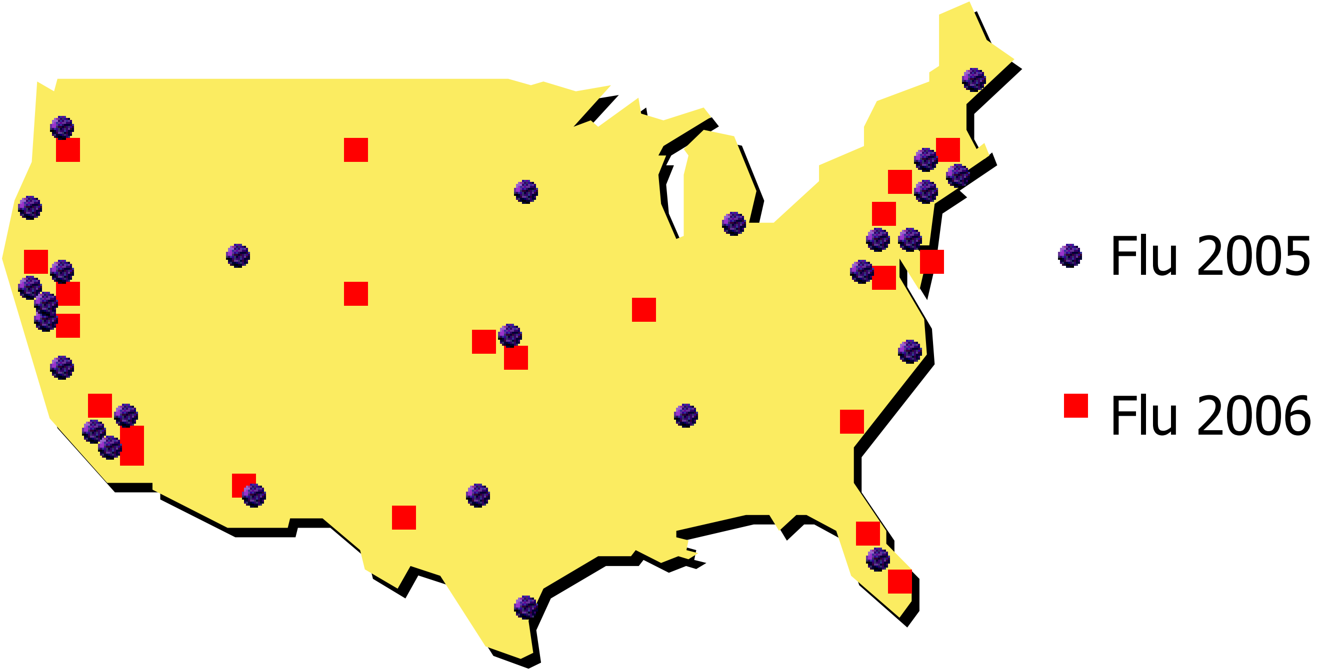
Transactions of 30-40 yr olds



Outbreak of diseases

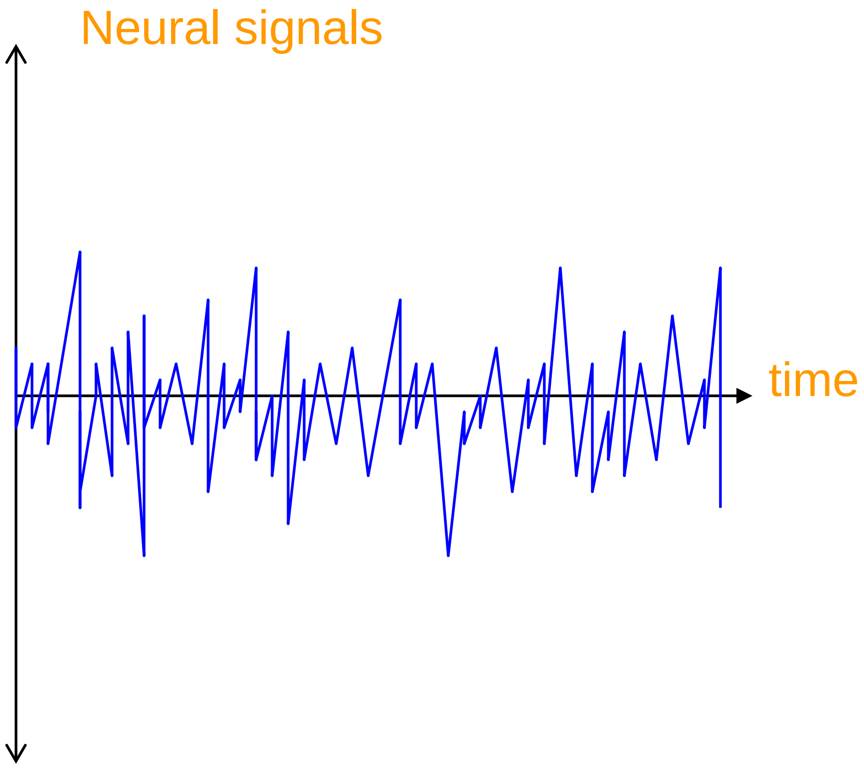


- Similar patterns?
- Correlated with income level?
- More prevalent near large airports?



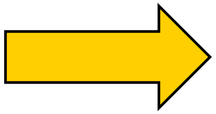
Information in neural spike trails

[Strong, Koberle, de Ruyter van Steveninck, Bialek '98]



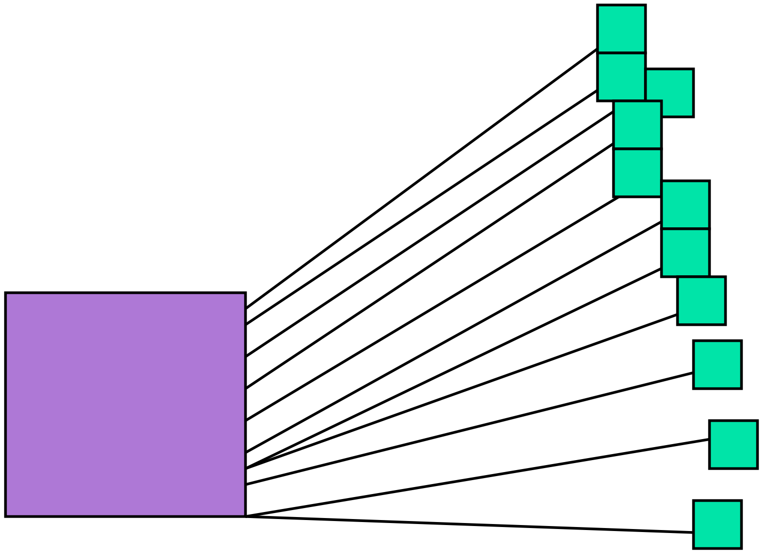
- Each application of stimuli gives sample of signal (spike trail)
- Entropy of (discretized) signal indicates which neurons respond to stimuli

Compressibility of data



Worm detection

- find ``heavy hitters’’ – nodes that send to many distinct addresses



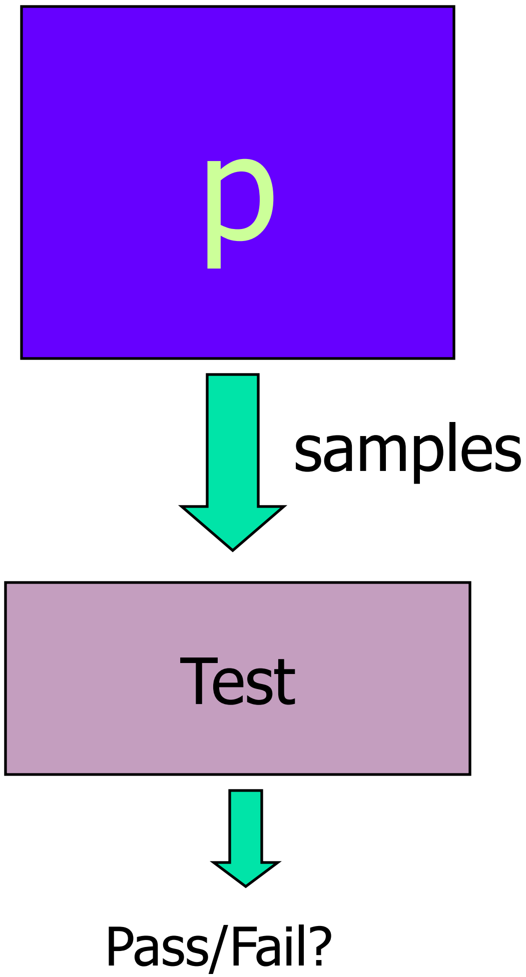
Testing properties of distributions:

- Decisions based on samples of distribution

- Focus on large domains
  - Can sample complexity be *sublinear* in size of the domain?

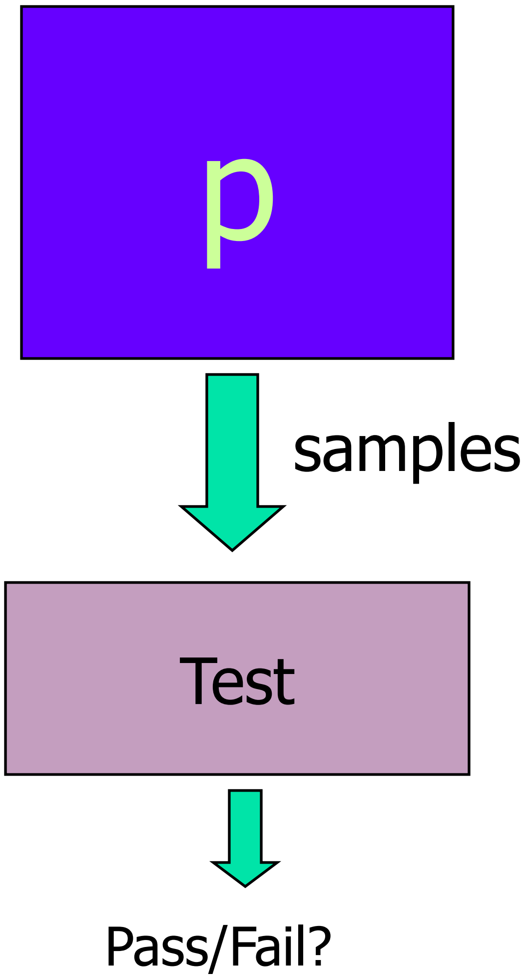
Rules out standard statistical techniques, learning distribution

Model:



- p is arbitrary black-box distribution over $[n]$, generates iid samples.
- p_i = Prob[p outputs i]
- Sample complexity in terms of n?

Is *p* uniform?



- Theorem: ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

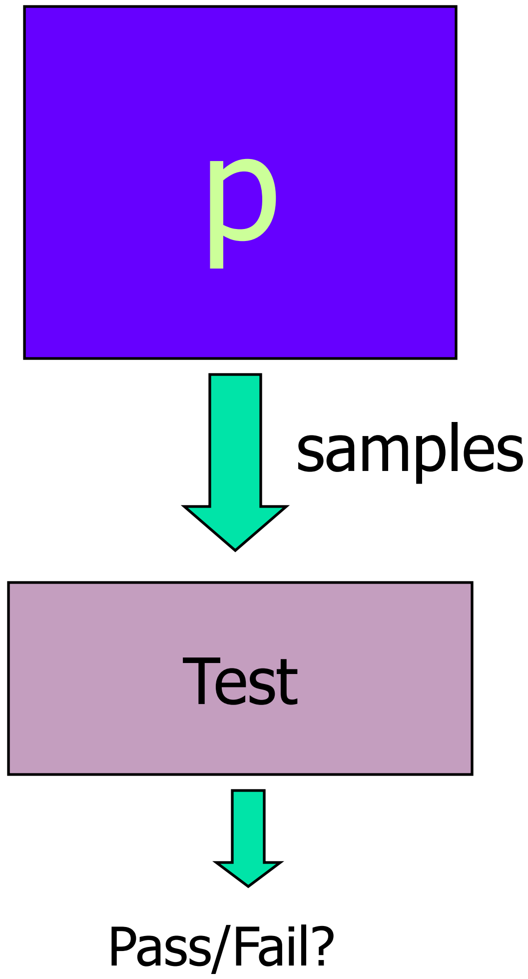
  $p=U$

  from $|p-U|_1>\varepsilon$ is $\theta(n^{1/2})$

  $|p-q|_1=\Sigma|p_i-q_i|$

- Nearly … test if p … distribution q [Batu Fischer Fortnow Kumar R. White]: "Testing identity"

Is p uniform?



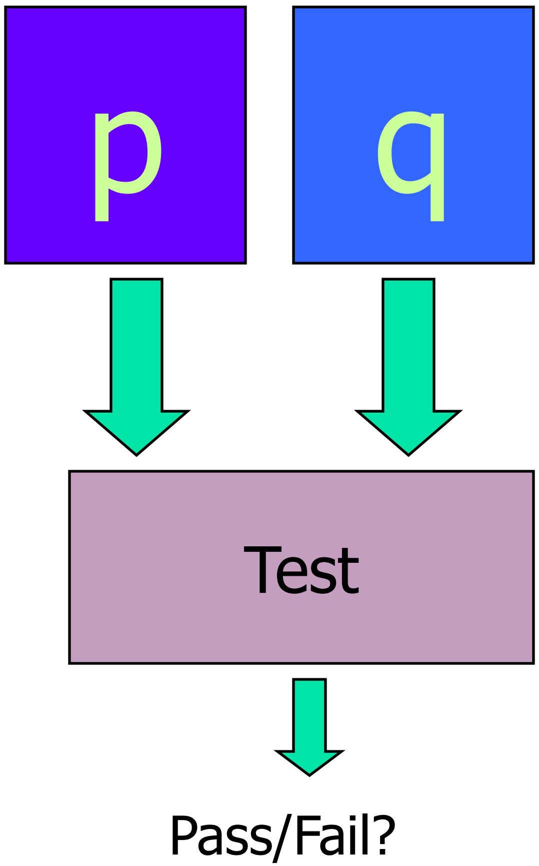
- Theorem: ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

  $$p=U$$

  from $|p-U|_1>\varepsilon$ is $\theta(n^{1/2})$

- Nearly same complexity to test if p is any *known* distribution q [Batu Fischer Fortnow Kumar R. White]: "Testing identity"

Testing closeness



Theorem: ([BFRSW] [P. Valiant])

Sample complexity of distinguishing

$p=q$

from $|p-q|_1 > \varepsilon$

is $\tilde{\theta}(n^{2/3})$

Approximating the distance between two distributions?

Distinguishing whether $|p\text{-}q|_1 < \varepsilon$ or $|p\ \text{-}q|_1$ is $\Theta(1)$ requires nearly linear samples [P. Valiant 08]

Some other properties (ignoring logs)

- Entropy estimation
  - *$\theta(n^{1/\gamma^2})$ for γ-approx* *[BDKR, V,BS, GKV]*
- Independence properties
  - Total independence of pairs [n]x[m]
    - *$O(n^{2/3}m^{1/3})$* *[BFFKRS]*
  - K-wise independence of binary N-vector
    - *$O(N^k)$, $\Omega(N^{k/2})$* *[AAKMRX]*
  - Almost k-wise independence
    - O(k log N), Ω(sqrt(k log N)) [AAKMRX]
- Monotonicity
  - $\theta(n^{1/2})$ for total order… [BKR],[BFRV]
- Support size
  - almost linear [RRSS]

Other properties to consider?

- Mixtures of *k* Gaussians
- “Junta”-distributions
- Clusterable-distributions
- Convex distributions
- ``Lipshitz’’ distributions
- Generated by a small Markovian process
- …

Getting past the lower bounds

- Special distributions
  - e.g, uniform on a subset, monotone
- Other query models
  - Queries to probabilities of elements
- Other distance measures

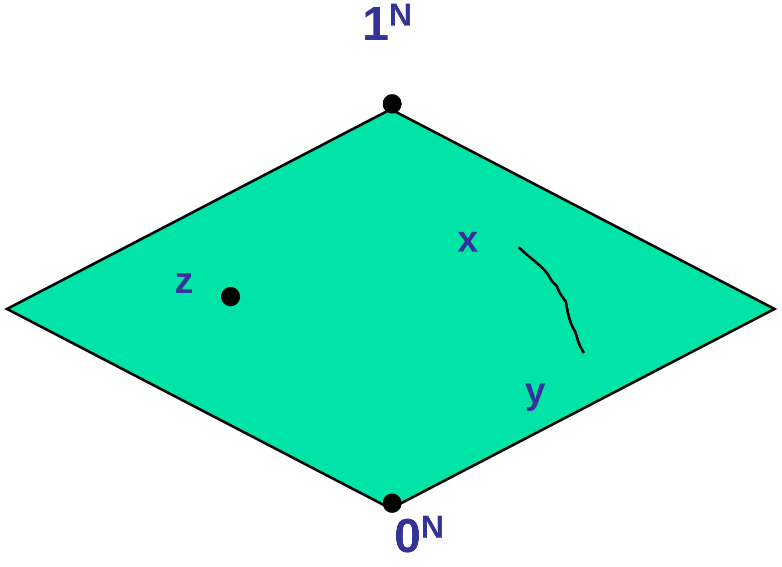
Monotone distributions over totally ordered domains

- Big wins:
  - Test uniformity with *O(1)* samples [Batu Kumar R.]
  - Other tasks doable with polylogarithmic samples: [Batu Dasgupta Kumar R.][BKR]
    - Testing closeness
    - Testing independence
    - Estimating entropy
- Do these big wins carry over to partial orders?

Monotone high-dimensional distributions

Domain: Boolean cube $\{0,1\}^N$

Are there testing algorithms with sample complexity polylogarithmic in domain size, i.e. *poly(N)*?



Testing Uniformity

Theorem [R. Servedio][Adamaszek Czumaj Sohler]**:** There is an $O(N/\varepsilon^2)$ sample complexity tester which given an unknown monotone distribution p over $\{0,1\}^N$ ($[0,1]^N$) satisfies (with probability 2/3):

- If $p=U$, algorithm outputs **"uniform"**
- If $\|p - U\|_1 > \varepsilon$, algorithm outputs **"far from uniform"**

- Comment: Nearly best possible

Bad news for Boolean cube

[R. Servedio]

- Technique for sample complexity lower bounds: monotone subcube decomposition
  - $2^{\Omega(N)}$ lower bound for testing equivalence to a known distribution (even product distributions!)
  - $2^{\Omega(N)}$ lower bound for approximating entropy

Open question for Boolean cube

Can one test monotone distributions over $\{0,1\}^N$ for any of the following properties

- equivalence to a known distribution
- approximating entropy
- independence

with fewer samples than for arbitrary distributions?

What about other partial orders?

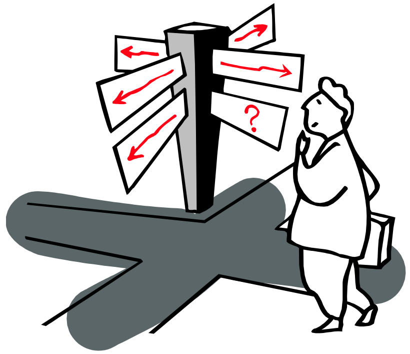
Other query models:

- Distribution given explicitly [BDKR]
- Distribution given both by samples and oracle for p_i's [BDKR][RS]
  - Can estimate entropy in polylog(n) time

Other distance measures:

- Earth Mover Distance [Doba Nguyen[2] R.]
  - Measures min weight matching to some distribution with the property
  - Can estimate distance between distributions, independence over $[0,1]^N$, in time *independent* of domain size
  - Still exponential in N
    - Can improve for highly clusterable distributions

Conclusions and Future Directions



- Distribution property testing problems are everywhere
- Several useful techniques known
- Other properties for which sublinear tests exist?
- Special classes of distributions?
- Time vs. query complexity
- Other query models?
- Non-iid samples?

Thank you