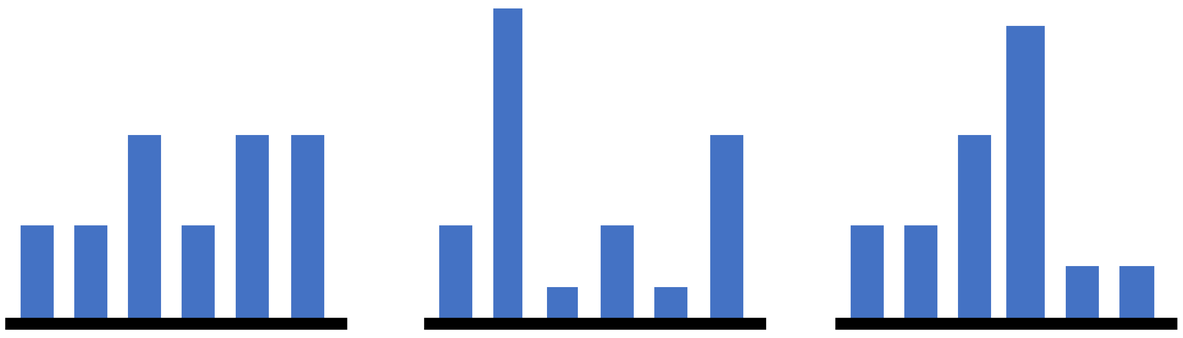
Statistical-Computational Trade-offs for Hypothesis Selection

Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, Sandeep Silwal, Haike Xu

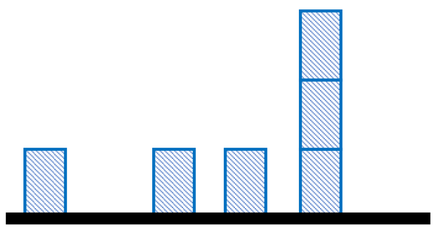
Searching over Distributions

Preprocess: k discrete distributions over $[n]$

$p_1, p_2, \ldots, p_k$



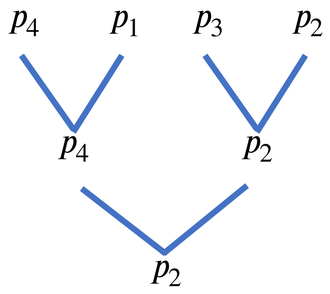
Query: samples from unknown distribution p_i



Goal: output p_j that is ϵ-close to p_i (TV distance)

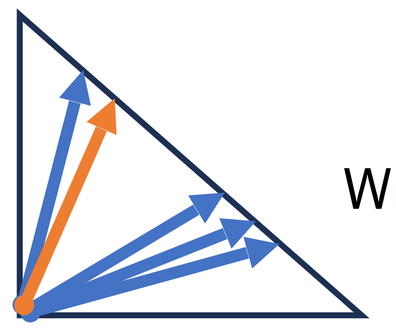
Nearest neighbor search over distributions

Optimizing Samples or Time



Samples: $O(\log k/\epsilon^2)$

Existing solutions use $\Omega(k)$ time (e.g., via a tournament)



Time: $O(k^\rho)$ query time

With $O(n/\epsilon^2)$ samples and $O(k^{1+\rho})$ space via LSH

Trivial solutions if we allow superpolynomial preprocessing/space

Sublinear samples (in n) and query time (in k)?

Our Results: Novel Statistical-Computational Trade-offs

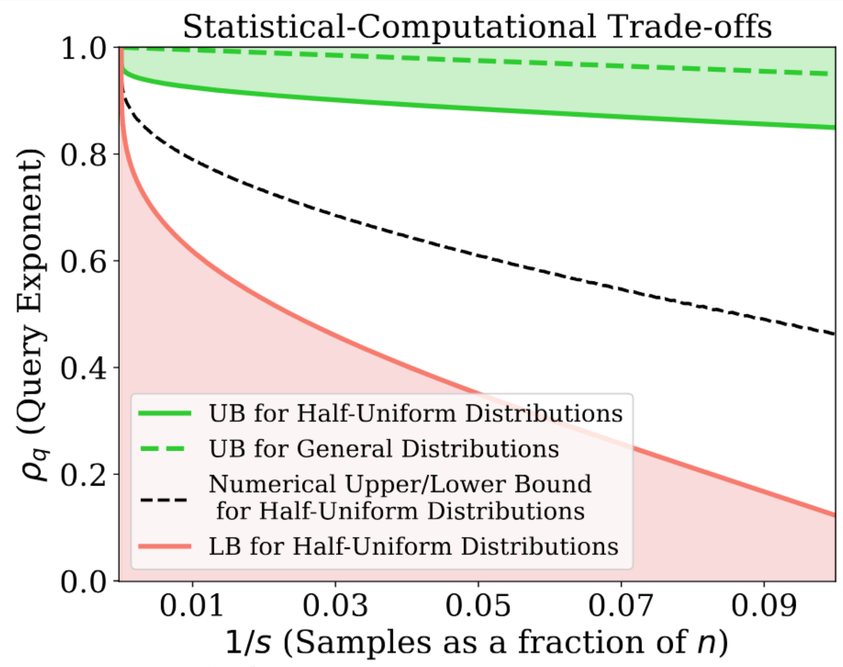
UPPER BOUND [AACINS'23]: Doubly sublinear data structure using hierarchy of LSH-based data structures on "heavy"/"light" elements.

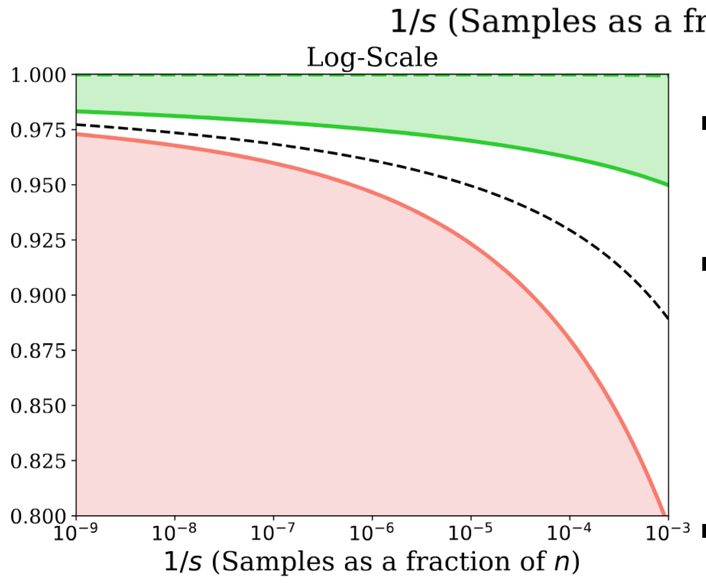
Samples	$O\left(\frac{n}{(\log k)^{1/4}\epsilon^2}\right)$
Runtime	$\tilde{O}\left(k^{1-1/(\log k)^{1/4}} + n\right)$
Preprocessing	Polynomial

- Insufficient samples to ***learn*** unknown p_i
- Insufficient runtime to ***scan*** over all distributions
- But both facts above are ***barely*** true!

Can we get truly sublinear samples and query time? $n^{0.9}$ samples and $k^{0.9}$ time?

LOWER BOUND [AACINSX'24]: Any data structure in the list of points model with $O\left(\frac{n}{\log(k)^{O(1)}}\right)$ samples must have runtime $k^{1-1/\log\log k}$.





- We build on and extend [AK '20]
- We give matching upper bound for hard instance of the main theorem
- Many open problems!

[AK '20]: Thomas D Ahle and Jakob BT Knudsen. Subsets and supermajorities: Optimal hashing-based set similarity search. FOCS '20.
[AACINS '23]: Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, Sandeep Silwal. Data Structures for Density Estimation. ICML'23.
[AACINSX '24]: Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, Sandeep Silwal, Haike Xu. Statistical-Computational Trade-offs for Density Estimation. NeurIPS'24.