Lecture 13

TTIC 41000: Algorithms for Massive Data

Toyota Technological Institute at Chicago

Spring 2021

Instructor: Sepideh Mahabadi

This Lecture

- Testing properties of distributions

Sublinear Time Algorithms

- The input is so huge that even reading all of it is not feasible
- Solve the problem accessing a *small* portion of the input
  - Need to specify the access model: what queries can be asked?
    - Random Access
      - E.g., For an array, given i, return the ith entry of a matrix, i.e., A[i]
      - For a graph, query the adjacency graph: given u,v, return A[u][v], i.e., does there exist an edge between u and v
      - Adjacency List: given u, i, return the ith neighbor of the vertex u (or Null if deg(u)<i)
    - Sample
      - Algorithm receives a random sample from a specific distribution
  - Parameters of interest
    - Number of queries asked
    - Actual runtime (could be sublinear, polynomial, or even exponential)

Model

- There is an unknown distribution p over a domain of size $[n]$
  - We can receive iid samples from p
  - Let p_i be the probability of outputting i
- Interested to know if p has a property or far from having the property
  - E.g. being uniform
  - Being close to another distribution q
  - Monotonicity, Unimodal, k-modal, k-flat, …
- Need to specify the distance measure, i.e., L_1 or L_2, or KL-divergence, …
- Sublinear number of samples in n?

Testing Uniformity

Is a lottery fair?

Problem Definition

- There is an unknown distribution p over a domain of size $[n]$
  - We can receive iid samples from p
  - Let p_i be the probability of outputting i
- Goal:
  - pass uniform distribution
  - Fail distributions that are ϵ-far from uniform
    - L_1 distance: $\|p - U\|_1 = \sum_i |p_i - \frac{1}{n}| > \epsilon$
    - L_2 distance: $\|p - U\|_2^2 = \sum_i \left(p_i - \frac{1}{n}\right)^2 > \epsilon^2$
- Sample complexity in terms of n and ϵ?

Naïve approach

- Take m samples
- Compute the empirical distribution p', i.e., $p_i' = (\#\text{times i apprears})/m$
- If $\|p' - U\|_1 > \epsilon$ fail
- Otherwise pass

- Problem: need $\Omega(n)$ samples for this to work using Chernoff

Estimation in L_2 distance using Collision probability

- What is the probability of collision for two samples?
  - $\Pr_{s,t \in p}[s = t] = \sum_{a \in [n]} p(a)^2 = \|p\|_2^2$
- What is the collision probability of U?
  - $1/n$
- Algorithm: approximate collision probability and compare to $1/n$
  - $\|p - U\|_2^2 = \sum_{a \in [n]} \left(p(a) - \frac{1}{n}\right)^2 = \sum_{a \in [n]} p(a)^2 - (2/n) \sum_a p(a) + \sum_a \frac{1}{n^2}$
  - $= \sum_a p(a)^2 - \frac{2}{n} + \frac{1}{n} = \|p\|_2^2 - \frac{1}{n}$
- Sufficient to get an additive $\frac{\epsilon^2}{2}$ error for L_2^2
  - If $p = U$, then $\|p\|_2^2$ is $1/n$
  - If $\|p - U\|_2 > \epsilon$ then $\|p\|_2^2 > \frac{1}{n} + \epsilon^2$
  - So let the threshold for deciding be $\frac{1}{n} + \frac{\epsilon^2}{2}$

How many samples? How to use samples?

- Naïve idea: Take $\mathbf{2s}$ samples and count the number of collisions between every consecutive pair.
  - The pairs are independent
- More efficiently: take $\boldsymbol{s}$ samples and compare the collision between "all" pairs
  - Have some dependence now
  - Use variance to bound accuracy

Algorithm

- Take s samples $X_1, \cdots, X_s$
- For $1 \leq i < j \leq s$, let $\sigma_{i,j}$ be 1 if $X_i = X_j$ and 0 otherwise
- Output $A = \frac{\sum_{i<j} \sigma_{i,j}}{\binom{s}{2}}$

Need to show

- It works in expectation
- It works with good probability

Analyzing the expectation

- $\mathbb{E}[A] = \frac{\binom{s}{2}\mathbb{E}[\sigma_{i,j}]}{\binom{s}{2}} = \Pr[\sigma_{i,j} = 1] = \|p\|_2^2$
- Chebyshev $\Pr[|A - \mathbb{E}[A]| > \rho] \leq Var[A]/\rho^2$
- For additive approximation set $\rho = \epsilon$
- For multiplicative approximation set $\rho = \epsilon\|p\|_2^2$
- Bound $Var[A]$ and show that $\frac{Var[A]}{\epsilon^2\|p\|_2^4} \ll 1$ if $s = \Omega\left(\frac{\sqrt{n}}{\epsilon^2}\right)$
- Better bound is possible if we have a bound on the max prob of any element

Bounding the variance

Lemma: $Var\left[\sum_{i,j} \sigma_{i,j}\right] \leq 2\left(\binom{s}{2} \cdot \|p\|_2^2\right)^{\frac{3}{2}}$

- $\bar{\sigma}_{i,j} = \sigma_{i,j} - \mathbb{E}[\sigma_{i,j}]$
- $Var\left[\sum_{i,j} \sigma_{i,j}\right] = \mathbb{E}\left[\left(\sum_{i,j} \bar{\sigma}_{i,j}\right)^2\right] = \mathbb{E}\left[\sum_{i<j} \bar{\sigma}_{i,j}^2 + \sum_{i<j,k<l} \bar{\sigma}_{i,j}\bar{\sigma}_{k,l} + \sum_{i<j,i<l} \bar{\sigma}_{i,j}\bar{\sigma}_{i,l} + \sum_{i<j,k<j} \bar{\sigma}_{i,j}\bar{\sigma}_{k,j}\right]$
- $\mathbb{E}\left[\sum_{i<j} \bar{\sigma}_{i,j}^2\right] \leq \mathbb{E}\left[\sum_{i<j} \sigma_{i,j}^2\right] = \binom{s}{2} \cdot \|p\|_2^2$
- $\mathbb{E}\left[\sum_{i<j,k<l} \bar{\sigma}_{i,j}\bar{\sigma}_{k,l}\right] = \sum_{i,j,k,l} \mathbb{E}[\bar{\sigma}_{i,j}]\mathbb{E}[\bar{\sigma}_{k,l}] = 0$ by independence of samples.
- $\mathbb{E}\left[\sum_{i<j,i<l} \bar{\sigma}_{i,j}\bar{\sigma}_{i,l}\right] \leq \mathbb{E}\left[\sum_{i,j,l} \sigma_{i,j}\sigma_{i,l}\right] \leq \binom{s}{3} \sum_x p(x)^3 \leq \frac{s^3}{6}\|p\|_3^3 \leq \frac{\sqrt{3}}{2}\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$
- $Var\left[\sum_{i,j} \sigma_{i,j}\right] \leq \binom{s}{2} \cdot \|p\|_2^2 + 0 + \sqrt{3}\left(\binom{s}{2}\|p\|_2^2\right)^{\frac{3}{2}} \leq 2\left(\binom{s}{2}\|p\|_2^2\right)^{3/2}$
- $\frac{Var[A]}{\epsilon^2\|p\|_2^4} \leq \frac{2\left(\binom{s}{2}\|p\|_2^2\right)^{\frac{3}{2}} \cdot \frac{1}{\binom{s}{2}}}{\epsilon^2\|p\|_2^4} \leq 2\binom{s}{2}^{-\frac{1}{2}}\|p\|_2^{-1}\epsilon^{-2} \leq 1/3$ if $s = \Omega(\frac{\sqrt{n}}{\epsilon^2})$

Overview of other properties

Closeness of two distributions

- Algorithm knows q and wants to realize if p and q are close or far.

- Reduction to uniformity testing
  - Relabel the domain so that q is monotone (we know q) so this can be done
  - Partition the domain into $O(\log n)$ parts, so that each group is almost flat
    - Differ by $(1+\epsilon)$ multiplicative
    - q is close to uniform in each part
  - Test
    - p is close to uniform in each part
    - p has the right weight in each bucket

Bucketing

- $R_0 = \left\{j: q(j) < \frac{1}{n \log n}\right\}$
  - Total probability of them is only $1/\log n$ which is less than ϵ
- $R_i = \left\{j: \frac{(1+\epsilon)^{i-1}}{n \log n} \leq q(j) < \frac{(1+\epsilon)^i}{n \log n}\right\}$
  - All probabilities are within a $(1+\epsilon)$ factor of each other
  - Total number of buckets is only $\frac{\log n}{\epsilon}$
- Let Z be the following distribution
  - Pick bucket i with probability $\sum_{j \in R_i} q(j)$
  - Pick an element uniformly at random from bucket i
- We show that Z and q are close

Single bucket

- Let
  - q_i be q conditioned on i-th bucket
  - U_i be uniform on the bucket
  - ℓ the number of elements in the bucket
- Lemma: q_i and U_i are ϵ −close under L_1 distance and ϵ^2/ℓ-close over L_2^2 distance
  - Let $x_1, \cdots, x_\ell$ be the conditional probabilities
  - Clearly, $x_1 \leq \frac{1}{\ell} \leq x_\ell$ and so $x_\ell \leq (1+\epsilon)x_1 \leq (1+\epsilon)/\ell$ and $x_1 \geq \frac{1}{\ell(1+\epsilon)} \geq \frac{1-\epsilon}{\ell}$
  - So $\left|x_j - \frac{1}{\ell}\right| \leq \epsilon/\ell$ and thus the L_1 distance is at most ϵ and the L_2^2 is at most $\frac{\epsilon^2}{\ell}$
  - So $\|q_i\|_2^2 \leq \left(1+\epsilon^2\right)/\ell$

Single bucket algorithm

- Algorithm: Estimate $\|p_i\|_2^2$ and fail if $> \frac{1+\epsilon^2}{|R_i|}$
- Lemma: if $\|p_i\|_2^2 \leq (1+\epsilon^2)/|R_i|$ then $\|q_i - p_i\|_1 \leq 2\epsilon$
  - Both q_i and p_i are close to uniform
  - Use triangle inequality

Overall algorithm

- Bucket q
- Calculate total weight of q in each bucket
- Estimate total weight p assigns to each bucket ($O(\log n)$ samples)
- If L_1 distance between bucket weights is more than ϵ, reject
- For each bucket with weight more than $\epsilon/2k$ where k is the number of buckets
  - Estimate collision probability p_i (need $O(\frac{\sqrt{n}k \log n}{\epsilon^2})$ samples of p)
  - Fail if the estimate is bigger than $(1+\epsilon^2)/|R_i|$

Correctness

- One way is clear
- If p and q pass the test
  - Total weight of skipped buckets is at most ϵ
  - p_i is ϵ-close to q_i in each bucket
  - Bucket weight of p and q are ϵ-close
- Overall they will be $O(\epsilon) - close$
- Testing identity can be reduced to $O(\log n)$ uniformity testing

Other properties

- ❑ Testing closeness: both q and p are unknown and we can get samples from them, requires $\Theta\left(n^{\frac{2}{3}}\right)$
  - Two phase approach:
    - Sample to detect heavy elements of both
    - Estimate distance of heavy elements and light elements separately
- ❑ Approximating distance between two distributions (if $\|p-q\|_1 < \epsilon$ or $\Omega(1)$) requires nearly linear samples)
  - Estimating $\|p-q\|_1$ requires $\Theta\left(\frac{n}{\log n}\right)$ samples.
- ❑ Testing independence where we receive samples from the joint distribution over [n]x[m], the goal is to check if the marginal are independent
  - Can be done in $\tilde{O}(n^{\frac{2}{3}}m^{\frac{1}{3}})$ assuming $n > m$
- ❑ …