COMS E6998-9: Algorithms for Massive Data (Fall'23) Oct. 25, 2023

Lecture 15

Instructor: *Alex Andoni* Scribe: *Krish Singal*

1 Distribution Testing

We continue our discussion of uniformity testing from last class.

1.1 Uniformity Testing

Algorithm 1 Uniformity Testing via Collision Count

Input: Samples $x_1, ..., x_m \sim \mathcal{D}$
$C \leftarrow |\{i < j | x_i = x_j\}|$
$M \leftarrow \binom{m}{2}$
if $\frac{C}{M} < \frac{1+\frac{\epsilon^2}{2}}{n}$ **then**
  return Uniform
else
  return ϵ-far from uniform

Let $d \triangleq \|\mathcal{D}\|_2^2$. Last class, we proved

1. $\mathcal{D} = U_n \implies \mathbb{E}[\frac{C}{M}] = \frac{1}{n}$
2. $\mathcal{D}$ ϵ-far from $U_n \implies \mathbb{E}[\frac{C}{M}] = \|\mathcal{D}\|_2^2 \geq \frac{1+\epsilon^2}{n}$

We claim that the collision rate $\frac{C}{M}$ concentrates around d for well chosen number of samples m.

Claim 1. $\Pr[|\frac{C}{M} - d| > \frac{\epsilon^2}{3} d] \leq 0.1$ *for* $m = O(\frac{\sqrt{n}}{\epsilon^4})$.

Proof. We compute the variance of C towards applying Chebyshev's inequality

$$\begin{aligned}
\mathrm{Var}[C] = \mathbb{E}[C^2] - \mathbb{E}[C]^2 &= \mathbb{E}\left[\left(\sum_{i<j} \mathbf{1}[x_i = x_j]\right)^2\right] - [Md]^2 \\
&= \sum_{i<j}\sum_{i'<j'} \mathbb{E}\left[\mathbf{1}[x_i = x_j]\mathbf{1}[x_{i'} = x_{j'}]\right] - [Md]^2 \\
&\leq \mathbb{E}[C]^2 + \sum_{i<j} \Pr[x_i = x_j] + 2\sum_{i<j, j' \neq i,j} \mathbb{E}[\mathbf{1}[x_i = x_j = x_{j'}]] - [Md]^2 \\
&= \sum_{i<j} \Pr[x_i = x_j] + 2\sum_{i<j, j' \neq i,j} \mathbb{E}[\mathbf{1}[x_i = x_j = x_{j'}]]
\end{aligned}$$

Observe that the uniform distribution U_n is the unique minimizer of $\min_{\mathcal{D}} \|D\|_2^2$. Thus, $d \geq \frac{1}{n}$.

$$\begin{aligned}
&\leq Md + 2m^3\|\mathcal{D}\|_3^3 \\
&\leq Md^2n + 2m^3d^{3/2} && (d \geq \tfrac{1}{n} \text{ and } \|\cdot\|_2 \geq \|\cdot\|_3) \\
&\leq \frac{n}{M} \cdot M^2d^2 + 2m^3d^{3/2} \cdot \sqrt{dn} && (d \geq \tfrac{1}{n}) \\
&\leq M^2d^2(\frac{n}{M} + 8\frac{\sqrt{n}}{m}) \\
&\leq M^2d^2 \cdot 9\frac{\sqrt{n}}{m}
\end{aligned}$$

To apply Chebyshev's, we need

$$\text{Var}[C] \leq \frac{1}{10}(\frac{\epsilon^2}{3}dM)^2$$

So,

$$\begin{aligned}
M^2d^2 \cdot 9\frac{\sqrt{n}}{m} &\leq \frac{1}{10} \cdot \frac{\epsilon^4}{9} \cdot d^2M^2 \\
\implies m &\geq 810\frac{\sqrt{n}}{\epsilon^4} = \Theta(\frac{\sqrt{n}}{\epsilon^4})
\end{aligned}$$

$m = \Theta(\frac{\sqrt{n}}{\epsilon^4})$ suffices. □

Thus, uniformity testing via collision counting gives the guarantees that

1. If $\mathcal{D} = U_n$, then with probability ≥ 0.9

$$\frac{C}{M} \leq \frac{1}{n} + \frac{\epsilon^2}{3} \cdot \frac{1}{n} = \frac{1 + \epsilon^2/3}{n}$$

in which case we accept.

2. If $\mathcal{D}$ is ϵ-far from uniform, then with probability ≥ 0.9

$$\begin{aligned}
\frac{C}{M} &\geq d - \frac{\epsilon^3}{3}d = d(1 - \frac{\epsilon^2}{3}) \\
&\geq \frac{1}{n}(1 - \frac{\epsilon^2}{3})(1 + \epsilon^2) \\
&= \frac{1}{n}(1 - \frac{\epsilon^2}{3} + \epsilon^2 - \frac{\epsilon^4}{3}) \\
&\geq \frac{1}{n}(1 + \frac{\epsilon^2}{2}) && (\text{for } \epsilon \text{ small enough})
\end{aligned}$$

in which case, we reject.

1.2 Closeness Testing (with known $\mathcal{Q}$)

We now consider testing closeness between unknown distribution $\mathcal{D}$ and known distribution $\mathcal{Q}$. The task is to distinguish between (1) $\mathcal{D} = \mathcal{Q}$ and (2) $\mathcal{D}$ is ϵ-far from $\mathcal{Q}$.

Theorem 2. *There exists an $O(\sqrt{n} \cdot (\frac{1}{\epsilon})^{O(1)})$ closeness-tester*

Proof. We only prove the theorem for the special case $\forall i. Q_i \in \frac{1}{n} \cdot \mathbb{N}$.

We map closeness testing over $[n]$ to uniformity testing over a new domain S, where $|S| = O(n)$. We define $s_i = n \cdot \mathcal{Q}_i$ and flatten distribution $\mathcal{Q}$ to $\mathcal{Q}'$ which is uniform over

$$S = \bigcup_{\substack{i=1 \\ s_i \neq 0}}^{n} i \times \{1, 2, ..., s_i\}$$

namely $\mathcal{D}'_{(i,j)} = \frac{\mathcal{D}_i}{s_i}$. Notice that $\mathcal{D} = \mathcal{Q} \implies \mathcal{D}' = \mathcal{Q}'$. We also claim that $\|\mathcal{D}' - \mathcal{Q}'\|_1 = \|\mathcal{D} - \mathcal{Q}\|_1$. We show it directly from the definition of $\mathcal{D}'$

$$\|\mathcal{D}' - \mathcal{Q}'\|_1 = \sum_i \sum_{j=1}^{s_i} |\frac{\mathcal{D}_i}{s_i} - \frac{\mathcal{Q}_i}{s_i}| = \sum_i |\mathcal{D}_i - \mathcal{Q}_i| = \|\mathcal{D} - \mathcal{Q}\|_1$$

Then, we can do uniformity testing of $\mathcal{D}'$ over S (reject if any sample $x = i$ such that $s_i = 0$). Thus, the sample complexity is $m = O_\epsilon(\sqrt{|S|}) = O_\epsilon(\sqrt{n})$. □

Theorem 2 shows that $O_\epsilon(\sqrt{n})$ is optimal for general $\mathcal{Q}$, but for distributions $\mathcal{Q}$ with special structure, we might be able to do better. [1] takes advantage of $\mathcal{Q}$ with special structure and gives improved sample complexity bounds. It uses the quantity

$$\sum_i \frac{(m\widehat{\mathcal{D}}_i - m\mathcal{Q}_i)^2 - m\widehat{\mathcal{D}}_i}{\widehat{\mathcal{D}}_i^{2/3}}$$

to determine whether to accept or reject. This is very similar to the χ^2-test by Pearson in 1900 which uses the quantity

$$\sum_i \frac{(m\widehat{\mathcal{D}}_i - m\mathcal{Q}_i)^2 - m\widehat{\mathcal{Q}}_i}{\mathcal{Q}_i}$$

1.3 Other Problems

1. **Closeness Testing (with unknown $\mathcal{Q}$):** We are given sample access to $\mathcal{Q}$ and $\mathcal{D}$ – both unknown distributions. The optimal sample complexity in this setting is known to be $\Theta(n^{2/3})$

2. **Independence Testing:** We are given sample access to $\mathcal{D}$ over $[n] \times [n]$. The task is to determine whether the marginal distributions are independent or ϵ-far from independent.

3. **Tolerant Testing:** A different model of property testing where we wish to distinguish whether $\mathcal{D}$ is ϵ_1 close to some property $\mathcal{P}$ or ϵ_2-far.

2 Sublinear Time Algorithms

2.1 Monotonocity Testing

We are given query access to a string $x \in \mathbb{N}^n$, and we want to answer whether x is increasing. We say that x is ϵ-far from increasing if deleting ϵn entries of x cannot make it increasing (equivalently if $LIS(x) < (1-\epsilon)n$).

Theorem 3. *There exists a one-sided monotonicity tester that takes $O(\frac{\log n}{\epsilon})$ time.*

Before proving the theorem, we explore two potential ideas. We naturally first consider drawing random indices $i < j$ and checking whether $x_i < x_j$. An adversarial case such as $x = 2, 1, 4, 3, 6, 5, ...$ only has $\approx \frac{n}{2}$ violating pairs, so $\Theta(n)$ draws are needed in expectation to find one. To remedy performance on cases such as this where violations are localized, we consider drawing random index i and checking whether $x_i < x_{i+1}$. However, we quickly notice that another adversarial case $x = \frac{n}{2}, \frac{n}{2}+1, ..., n, 1, 2, ..., \frac{n}{2}-1$ has only one violating index, so once again $\Theta(n)$ draws are needed in expectation to find it.

To capture the possibilities of both local and global violations, we try taking pairs i, j at distances 2^k for all $k \in [\log n]$ from one another. Consider the following algorithm

Algorithm 2 Monotonicity Testing

```
for iter = 1, ..., T = O(1/ε) do
    Let i ∈_r [n]
    Binary search for y ≜ x_i in x[1, ..., n] to get index j
    if j ≠ i then
        return Reject
return Accept
```

with the following binary search subroutine

Algorithm 3 Binary Search(y)

```
Input: Interval [s, t]
m ← ⌊(s+t)/2⌋
if x_m = y then
    return m
if x_m < x_s or x_m > x_t then
    return Reject
if y < x_m then
    Recurse on [s, m]
else
    Recurse on [m, t]
```

Claim 4. *If x is ϵ-far from increasing, then $\Pr_{i \in_r [n]}[\textit{Binary Search Fails}] \geq \epsilon$.*

We will prove correctness in the next class.

References

[1] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In Proceedings of 25th ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1193–1203, 2014. arXiv:1308.3946.