**ECS 189A Sublinear Algorithms for Big Data** **Fall 2024**

Lecture 11: Uniformity testing + Lower bound of Omega(sqrt(n))

*Lecturer: Jasper Lee* *Scribe: Ben Wiesner*

# 1 Uniform Testing

Consider $\mathcal{C}$ being the set of all distributions on $[n]$, and let $\mathcal{P}$ be the singleton set containing the uniform distribution on $[n]$.
Given i.i.d. samples from a distribution $p$, we aim to test:

$$p \in \mathcal{P} \quad \text{(i.e., } \mathcal{P} = \text{Unif}[n]) \quad \text{versus} \quad p \text{ is } \varepsilon\text{-far from Unif}[n]$$

in TV-distance with probability $\geq \frac{2}{3}$.
Goal: Minimize time and query sample complexity($m$).

**Remark.** In this lecture, we only focus on the constant probability regime. The high probability regime is much more complicated.

Most modern and best-known results on uniformity testing: Gupta and Price (2022).

# 2 Warm Up (Special Case)

We aim to distinguish between the distributions Unif$[2n]$ and Unif$(A)$, where:
$|A| = n$, and $A \subseteq [2n]$ is chosen adversarially.
Note that for any particular $A$:

$$d_{TV}\left(\text{Unif}[2n], \text{Unif}(A)\right) = \frac{1}{2}$$

- **Idea**: This involves a collision (birthday paradox) bound.
- **Birthday Paradox** refers to the counter intuitive fact that a group of 23 people have a 50 percent chance of sharing a birthday. For more information you can visit this Wikipedia article.

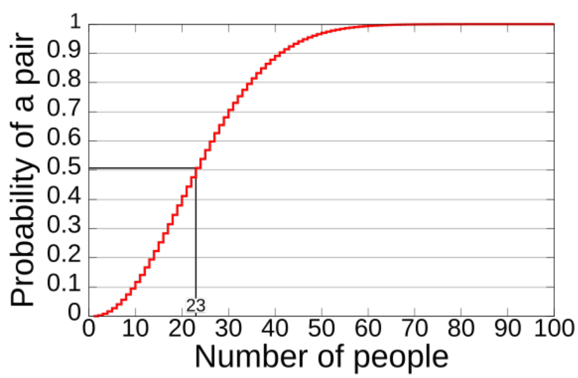


Figure 1: Birthday Problem

Consider a set $S$ of size $k$, and suppose we draw $m$ samples from Unif($S$). Then, the probability of not seeing a collision is:

$$\begin{aligned}\mathbb{P}(\text{no collision}) &= \prod_{i=1}^{m}\left(1-\frac{i-1}{k}\right)\\ &\le \prod_{i=1}^{m}\exp\left(-\frac{i-1}{k}\right) && (1+x\le e^x)\\ &= \exp\left(-\frac{1}{k}\sum_{i=1}^{m}(i-1)\right)\\ &= \exp\left(-\frac{1}{k}\cdot\frac{m(m-1)}{2}\right)\end{aligned}$$

We want to also lower bound the probability above. However, the reverse of our favorite inequality is clearly not true. However, $1-x \ge e^{-1.01x}$ for a sufficiently small positive $x$. Assuming $m$ is not too big relative to $k$ and $k$ is very large, then each $1-\frac{i-1}{k}$ in the product will be sufficiently small, and hence we can apply the latter inequality. Therefore, for $m \ll k$ and $k \gg 1$,

$$\begin{aligned}\mathbb{P}(\text{no collision}) &= \prod_{i=1}^{m}\left(1-\frac{i-1}{k}\right)\\ &\ge \prod_{i=1}^{m}\exp\left(-1.01\frac{i-1}{k}\right)\\ &= \exp\left(-\frac{1.01}{k}\cdot\frac{m(m-1)}{2}\right)\end{aligned}$$

Taking $m = O(\sqrt{n})$ samples creates a constant gap in the probability of observing a collision. Thus, by repeating $O(1)$ trials and using Chebyshev's inequality or Hoeffdings bound, we can estimate the collision probability accurately enough to distinguish between the two scenarios.

—

Summary of Results:

- **Uniformity Testing in General**: Requires $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples.
- **Collision Tester**: Indeed requires only $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples.
- **Note**: $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ sample bound for the collision tester needs far harder analysis, so today we will show a weaker $O\left(\frac{\sqrt{n}}{\varepsilon^4}\right)$ sample bound instead.

**Fact 11.1.**

*1.* $P(x=y) = \sum_i p_i^2 = ||p||_2^2$

*2.* $||p-U_n||_2^2 = \sum(p_i - \frac{1}{n})^2 = \sum(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2}) = \left(\sum p_i^2\right) - \frac{1}{n}$ *($U_n$ = Uniform Dist.)*

$||p-U_n||_2^2 \ge 0 \Rightarrow ||p||_2^2 = \frac{1}{n}$ *iff* $p = U_n$

3. $d_{TV}(p_1, U_n) = \frac{1}{2}||p - U_n||_1 \leq \frac{\sqrt{n}}{2}||p - U_n||_2$

$||p - U_n||_1 = \sum |p_i - \frac{1}{n}| = \sum |p_i - \frac{1}{n}| \cdot 1$
*(cauchy schwartz)*

$\leq \sqrt{\sum |p_i - \frac{1}{n}|^2 \cdot n} = \sqrt{n}\sqrt{||p - U_n||_2^2}$

**Corollary 11.2.**

*If $p$ is $\varepsilon$-far from $U_n$, then*

$$\|p\|_2^2 \geq \frac{1 + 4\varepsilon^2}{n}$$
$$\varepsilon \leq d_{TV}(p, U_n) \leq \frac{\sqrt{n}}{2}||p - U_n||_2 = \frac{\sqrt{n}}{2}\sqrt{||p||_2^2 - \frac{1}{n}}$$
$$\Rightarrow \frac{2\varepsilon}{n} \leq \sqrt{||p||_2^2 - \frac{1}{n}}$$

*Question: Using $\frac{1+4\varepsilon^2}{n}$ gap how do we estimate$\|p\|_2^2$ to good accuracy using $O(\frac{\sqrt{n}}{\varepsilon^4})$ samples?*

---

**Algorithm 11.3**

---

1. Take $m$ samples from $p$.
2. Compute $Y_{ij} = \mathbb{1}_{\{x_i = y_j\}}$, Compute $C = \sum_{i<j} \frac{Y_{ij}}{\binom{m}{2}}$.
3. Accept if $C \leq \frac{1+\varepsilon^2}{n}$.

Note: We want to show C is concentrated around the expectation. However, $C$ is not a sum of independent terms, so we will bound $\mathrm{Var}(C)$ and apply Chebyshevs inequality.

---

**Theorem 11.4.**
*Alg 11.3 on input $O(\frac{\sqrt{n}}{\varepsilon^4})$ samples, tests uniformity (vs $\varepsilon$-far) with probability $\geq \frac{2}{3}$*
***Proof***
*We know that $\mathbb{E}C = \|p\|_2^2$, now need to compute $\mathrm{Var}(C) = \mathbb{E}(C^2) - (\mathbb{E}(C))^2$.*
******Intuition******: In the uniform case, the test statistic will be centered around the expectation $1/n$. In any other case we know that the collision statistic $C$ will be centered around at least $\frac{1+4\varepsilon^2}{n}$. We need to control the overlap between these two collision-statistic distributions. In the non-uniform case, if the mean of the collision statistic is close to $\frac{1+4\varepsilon^2}{n}$ we can only afford a small variance for it, to separate it from the uniform case. If the mean is much larger than $\frac{1+4\varepsilon^2}{n}$ though, then we can afford a larger variance.*

$$
\begin{aligned}
\mathbb{E}(C^2) &= \binom{m}{2}^{-2} \mathbb{E}\left(\sum_{i<j} Y_{ij}^2 + \sum_{(i<j)\neq(k<l)} Y_{ij}Y_{kl}\right) \\
&= \binom{m}{2}^{-2} \|p\|_2^2 + \binom{m}{2}^{-2} \mathbb{E}\left(\sum_{\{i<j<k<l\}=3} Y_{ij}Y_{jk}\right) + \binom{m}{2}^{-2} \mathbb{E}\left(\sum_{\{i,j,k,l\}=4} Y_{ij}Y_{kl}\right) \\
&\leq \binom{m}{2}^{-2} \|p\|_2^2 + O\left(\binom{m}{2}^{-2}\binom{m}{3}\|p\|_3^3\right) + \underbrace{\binom{m}{2}^{-2}\binom{m}{2}\binom{m}{2}\|p\|_2^4}_{(\mathbb{E}(C))^2} \\
&= \binom{m}{2}^{-2} \|p\|_2^2 + \binom{m}{2}^{-2}\binom{m}{3}\|p\|_3^3 + (\mathbb{E}(C))^2
\end{aligned}
$$

*Therefore,*

$$
\mathrm{Var}(C) \leq \binom{m}{2}^{-2} \|p\|^{-2} + O\left(\frac{\|p\|_3^3}{m}\right)
$$

*Then,*

$$
\begin{aligned}
\mathbb{P}\left(\left|C - \|p\|_2^2\right| > \Theta(\varepsilon^2)\|p\|_2^2\right) &\leq O\left(\frac{\mathrm{Var}(C)}{\varepsilon^4\|p\|_2^4}\right) \\
&\leq O\left(\frac{1}{m^2\varepsilon^4\|p\|_2^2}\right) + O\left(\frac{\|p\|_3^3}{m\varepsilon^4\|p\|_4^2}\right)
\end{aligned}
$$

**Fact 11.5.** *$\|p\|_a \geq \|p\|_b$ for $a \leq b$ and $\|p\|_2^2 \geq \frac{1}{n}$.*

*Using this fact, we further deduce that*

$$
\mathbb{P}\left(\left|C - \|p\|_2^2\right| > \Theta(m^2)\|p\|_2^2\right) \leq \underbrace{O\left(\frac{n}{m^2\varepsilon^4}\right)}_{\text{if } n \gg \frac{\sqrt{n}}{\varepsilon^2}} + \underbrace{O\left(\frac{\sqrt{n}}{m\varepsilon^4}\right)}_{\text{if } m \gg \frac{\sqrt{n}}{\varepsilon^4}}
$$

*If $p$ = Unif[n], with probability $\geq \frac{2}{3}$*

$$
C \leq (1 + 0.1\varepsilon^2)\|p\|_2^2 = \frac{1 + .1\varepsilon^2}{n}
$$

*If $p$ is $\varepsilon$-far, with probability $\geq \frac{2}{3}$*

$$
\begin{aligned}
C &\geq (1 - 0.1\varepsilon^2)\|p\|_2^2 \\
&\geq (1 - 0.1\varepsilon^2)\left(\frac{1 + 4\varepsilon^2}{n}\right) \\
&\geq \frac{1 + 2\varepsilon^2}{n}
\end{aligned}
$$