CSCI 1951-W Sublinear Algorithms for Big Data **Fall 2020**

Lecture 13: Uniformity Testing

Lecturer: Jasper Lee *Scribe: Michael Nisenzon*

1 Problem Setting

Today, we will design and prove an algorithm to test whether a distribution is uniform or ϵ-far from uniform, with total variation distance as our metric. Intuitively, we are checking whether an n-sided coin is fair. In particular, we will show that the sample complexity of uniformity testing has an $O\left(\frac{\sqrt{n}}{\epsilon^4}\right)$ upper bound and a $\Omega(\sqrt{n})$ lower bound, although the same tester has a more refined analysis showing that its sample complexity is in fact $O\left(\frac{\sqrt{n}}{\epsilon^2}\right)$. A corresponding $\Omega\left(\frac{\sqrt{n}}{\epsilon^2}\right)$ lower bound can also be shown, but not covered in this lecture.

Let our property $P = \{\text{Unif}[n]\}$. We are given i.i.d. sample access to a distribution $\mathbf{p}$ over $[n]$, and we wish to distinguish with probability at least $\frac{2}{3}$ between the cases

$$\begin{cases} \mathbf{p} = \text{Unif}[n] \\ d_{\text{TV}}(\mathbf{p}, \text{Unif}[n]) > \epsilon \end{cases}$$

As a side note, the sample complexity in the high probability bound is substantially more complicated to analyse and show, which is why we focus on the constant probability regime in this lecture.

2 Building Intuition

Note that we can solve the problem with $O\left(\frac{n}{\epsilon^2}\right)$ samples by simply learning the distribution $\mathbf{p}$ to within total variation distance ϵ. However, we will show that we can do better.

2.1 Warmup

Consider the follow special case of uniformity testing, where we need to distinguish between $\text{Unif}[2n]$ and $\text{Unif}[A]$, where $|A| = n$ and $A \subset [2n]$ is chosen arbitrarily and unknown to the tester. We have that $d_{\text{TV}}(\text{Unif}[2n], \text{Unif}[A]) = \frac{1}{2}$, confirming that this problem is indeed a special case of uniformity testing, in the constant ϵ regime.

What is the right sample complexity for this problem? It turns out to be $O(\sqrt{n})$ (and in fact $\Omega(\sqrt{n})$ too as we will show at the end of the lecture). One way to achieve this is to consider *collisions* in the samples.

2.2 Collisions

We go back to the birthday problem, where we take m uniform and independent samples from a set S of size k and want to calculate/estimate the probability that no samples are equal to each other. By independence, we get

$$\mathbb{P}(\text{no collisions}) = \prod_{i=1}^{m} \left(1 - \frac{i-1}{k}\right)$$

Using the inequality $1 + x \leq e^x$, we can upper bound the probability by

$$\mathbb{P}(\text{no collisions}) \leq \exp\left(-\sum_{i=1}^{m} \frac{i-1}{k}\right) = \exp\left(-\frac{1}{k} * \frac{m(m-1)}{2}\right)$$

Further assuming that m/k is sufficiently small (in particular, in the regime where $m = \Theta(\sqrt{k})$), the inequality is a good approximation, say, within 1 percent. Concretely, for a sufficiently small $x > 0$, we have the inequality that $1 - x \geq e^{-1.01x}$. We then get a corresponding lower bound of

$$\mathbb{P}(\text{no collisions}) \geq \exp\left(-1.01\sum_{i=1}^{m} \frac{i-1}{k}\right) = \exp\left(-\frac{1.01}{k} * \frac{m(m-1)}{2}\right)$$

Taking $m = \Theta(\sqrt{n})$ for our warmup example, we have a constant gap between the probability of seeing collisions between the Unif$[n]$ and Unif$[2n]$ cases. As we have a constant gap, we can repeat the process a constant number of times to estimate the probability of seeing a collision using Hoeffding's inequality, to within some tiny constant additive error. We then check if our estimate is closer to the collision probability of Unif$[2n]$ or Unif$[A]$.

3 Upper Bound

We develop the previous idea further and give the *collision tester* for uniformity testing. First, we give a basic fact about the collision probability of a distribution.

Fact 13.1 *For an arbitrary distribution* $\mathbf{p}$ *over* $[n]$*, we have the following facts relating its collision probability to the* ℓ_2 *norm and distance from uniformity:*

$$\mathbb{P}_{x,y \leftarrow \mathbf{p}}(x = y) = \sum_i p_i^2 = ||\mathbf{p}||_2^2$$

$$||\mathbf{p} - \text{Unif}[n]||_2^2 = \sum_i (p_i - \frac{1}{n})^2 = \sum_i p_i^2 - \frac{1}{n} = ||\mathbf{p}||_2^2 - \frac{1}{n}$$

So $\mathbf{p} = \text{Unif}[n]$ *if and only if* $||\mathbf{p}||_2^2 = \frac{1}{n}$*. Furthermore, to relate the* ℓ_1 *and* ℓ_2 *distance from uniformity, we have*

$$d_{\text{TV}}(\mathbf{p}, \text{Unif}[n]) = \frac{1}{2}||\mathbf{p} - \text{Unif}[n]||_1 \leq \frac{\sqrt{n}}{2}||\mathbf{p} - \text{Unif}[n]||_2$$

where the inequality is just an application of the Cauchy-Schwarz inequality.

From the last item in Fact 13.1, we see that any lower bound ϵ for d_{TV} provides a lower bound for the L_2 distance. We capture this in a corollary:

Corollary 13.2 *If* $\mathbf{p}$ *is* ϵ*-far from* Unif$[n]$ *in* d_{TV}*, then* $||\mathbf{p}||_2^2 \geq \frac{1+4\epsilon^2}{n}$*.*

Corollary 13.2 is the structural property, in the form of gap in collision probability, that we will use to design and analyse our collision tester. The algorithmic question is how we can estimate $||\mathbf{p}||_2^2$ sufficiently accurately, using few samples.

We now state the collision tester.

Algorithm 13.3 (Collision Tester)

1. Take m samples $x_1, ..., x_m$ drawn i.i.d. from $\mathbf{p}$.
2. Compute $Y_{ij} = \mathbb{1}\{x_i = x_j\}$ and $C = \sum_{i<j} \frac{Y_{ij}}{\binom{m}{2}}$
3. Accept if and only if $C \leq \frac{1+0.01\epsilon^2}{n}$

We have that $C = \sum_{i<j} \frac{Y_{ij}}{\binom{m}{2}}$ is equal to the collision probability $||p||_2^2$ in expectation. Note however that C is *not* a sum of independent random variables, so we cannot apply any standard tail bound. Instead, we bound the variance of C and use Chebyshev's inequality to give constant probability concentration.

As mentioned before, our analysis will only give a sample complexity upper bound of $O(\sqrt{n}/\epsilon^4)$, even though a tight analysis shows a $O(\sqrt{n}/\epsilon^2)$ bound instead. The key difference between the two analyses lies in the tightness of bounding the variance.

We now state the sample complexity result.

Theorem 13.4 *Algorithm 13.3, on input $m = O\left(\frac{\sqrt{n}}{\epsilon^4}\right)$ samples tests uniformity vs ϵ-far from uniformity with probability at least $\frac{2}{3}$.*

Proof. As claimed above,

$$\mathbb{E}[C] = \sum_{i<j} \frac{\mathbb{E}[Y_{ij}]}{\binom{m}{2}} = ||\mathbf{p}||_2^2$$

To bound the variance, we bound the second moment of C:

$$\mathbb{E}[C^2] = \binom{m}{2}^{-2} \mathbb{E}\left[\sum_{i<j} Y_{ij}^2 + \sum_{(i<j)\neq(k<l)} Y_{ij}Y_{kl}\right]$$

The first term is the sum of indicator random variables, and the second can be split into two cases where either the (i, j) and (k, l) pairs share 1 index, or they have completely distinct indices. This gives

$$\mathbb{E}[C^2] = \binom{m}{2}^{-1} ||\mathbf{p}||_2^2 + \binom{m}{2}^{-2} \mathbb{E}\left[\sum_{\substack{|\{i,j,k,l\}|=3,\\(i<j),(k<l)}} Y_{ij}Y_{kl}\right] + \binom{m}{2}^{-2} \mathbb{E}\left[\sum_{\substack{|\{i,j,k,l\}|=4\\(i<j),(k<l)}} Y_{ij}Y_{kl}\right]$$

For the second term, $Y_{ij}Y_{kl}$ is non-zero if and only if the samples at all the three indices are equal, namely $x_i = x_j = x_k = x_l$ (noting that there are only 3 distinct indices in $\{i, j, k, l\}$). Each $Y_{ij}Y_{kl}$ term thus contributes an expectation of $||\mathbf{p}||_3^3$, and there are at most $\binom{m}{3}$ of these terms. For the third term, $Y_{ij}Y_{kl}$ is non-zero if and only if both Y_{ij} and Y_{kl} are non-zero, meaning that $x_i = x_j$ and $x_k = x_l$. Thus each $Y_{ij}Y_{kl}$ contributes an expectation of $||\mathbf{p}||_2^4$, and there are at most $\binom{m}{2}^2$ many of these terms. The above reasoning gives the following upper bound:

$$\begin{aligned}\mathbb{E}[C^2] \leq& \binom{m}{2}^{-1} ||\mathbf{p}||_2^2 + O\left(\binom{m}{2}^{-2}\binom{m}{3}||\mathbf{p}||_3^3\right) + \binom{m}{2}^{-2}\binom{m}{2}^2 ||\mathbf{p}||_2^4 \\ \leq& \binom{m}{2}^{-1} ||\mathbf{p}||_2^2 + O\left(\frac{||\mathbf{p}||_3^3}{m}\right) + ||\mathbf{p}||_2^4\end{aligned}$$

Finally, since $\mathbb{E}[C] = ||\mathbf{p}||_2^2$, we have

$$\mathrm{Var}(C) \leq \binom{m}{2}^{-1} ||\mathbf{p}||_2^2 + O\left(\frac{||\mathbf{p}||_3^3}{m}\right)$$

At this point, one potential worry is that the variance may be large, in particular when $||\mathbf{p}||_2^2$ is large ($\gg \frac{1}{n}$). However, this happens precisely when the distribution $\mathbf{p}$ is *far* from uniform, with a large gap in expectation, so it turns out to not be an issue.

Because of how the variance depends on the collision probability (which is the expectation), we apply Chebyshev's inequality to show that with high constant probability, we have a good *multiplicative* approximation to the expectation. This contrasts analysis in previous lectures where we typically asked for *additive* approximation to the expectation.

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

To further bound this probability, we need the following facts.

Fact 13.5

1. *From Fact 13.1, we know that* $||\mathbf{p}||_2^2 \geq \frac{1}{n}$.
2. *A standard fact about* ℓ_p *norms:* $||\mathbf{p}||_a \geq ||\mathbf{p}||_b$ *for* $1 \leq a \leq b$.

Applying Fact 13.5 to upper bound the first term, and to bound the ℓ_3 norm in the second term by the ℓ_2 norm, we have

$$\begin{aligned}\mathbb{P}(|C - ||\mathbf{p}||_2^2 > \Theta\left(\epsilon^2\right)||\mathbf{p}||_2^2) &\leq O\left(\frac{n}{m^2\epsilon^4}\right) + O\left(\frac{1}{m\epsilon^4||\mathbf{p}||_2}\right) \\ &\leq O\left(\frac{n}{m^2\epsilon^4}\right) + O\left(\frac{\sqrt{n}}{m\epsilon^4}\right)\end{aligned}$$

Taking $m = O\left(\frac{\sqrt{n}}{\epsilon^4}\right)$, we get that both terms are bounded above by any arbitrary constant, say they sum up to $\frac{1}{3}$.

Thus, we conclude that if $\mathbf{p} = \mathrm{Unif}[n]$, then $C \leq \frac{(1+0.01\epsilon^2)}{n}$ with probability at least $\frac{2}{3}$. Otherwise, if $\mathbf{p}$ is ϵ-far from $\mathrm{Unif}[n]$, then with probability at least $\frac{2}{3}$,

$$C \geq (1 - 0.01\epsilon^2)||\mathbf{p}||_2^2 \geq (1 - 0.01\epsilon^2)\frac{(1 + 4\epsilon^2)}{n} \geq \frac{(1 + 0.01\epsilon^2)}{n}$$

□

4 Lower Bound

We now wish to show a $\Omega(\sqrt{n})$ lower bound in the constant ϵ regime.

To show this, we revisit our warmup example. Consider again the task of distinguishing between $\mathrm{Unif}[2n]$ and $\mathrm{Unif}[A]$, where $A \subset [2n]$ and $|A| = n$. Suppose we take $m = c \cdot \sqrt{n}$ samples. For a sufficiently small constant c, we have that with probability at least 0.99

in both the Unif$[2n]$ and Unif$[A]$ cases, there are no collisions in the samples observed. This intuitively suggests that the two scenarios are behaving similarly and hence hard to distinguish. However, does the above actually give the desired indistinguishability result?

The key remaining issue is that we have not yet specified how we pick the set A. If A were a fixed set, then it is very easy to tell apart the two scenarios, simply by checking for samples in $[2n]/A$. On the other hand, if we pick A uniformly at random, then the distributions of samples are *exactly* the same in both cases *conditioned on seeing no collisions*, which happens with probability at least 0.99. Therefore, we have

$$d_{\mathrm{TV}}(\mathrm{Unif}^{\otimes m}[2n], \mathrm{Unif}^{\otimes m}[\text{random } A]) < \frac{1}{3}$$

for $m = c \cdot \sqrt{n}$ for a sufficiently small constant c.

Summarising, we have shown the following sample complexity lower bound.

Theorem 13.6 *Testing uniformity vs $\frac{1}{2}$-far from uniformity requires $\Omega(\sqrt{n})$ samples.*