1 Summary

We discuss the problem of determining whether a probability distribution is close to uniform, and begin to go over an algorithm for solving this with a sublinear number of samples.

2 Testing Uniformity

How can we tell if a probability distribution is uniform? We could try various statistical tests, such as $\chi^2$, Chernoff bounds, maximum likelihood, or learning the distribution.

Example: New Jersey Pick 3 / Pick 4 lottery

- Pick 3 — 8522 results, 1000 possibilities; $\chi^2$ test gives 42% confidence of uniformity
- Pick 4 — 6544 results, 10000 possibilities; $\chi^2$ test gives no confidence

The methods described above DON’T WORK in situations like Pick 4. They need too much data (all of them require roughly $\Theta(n)$ or $\Theta(n \log(n))$ samples to determine if a distribution over $\{1, \ldots, n\}$ is uniform.

Our goal will be to eventually construct a method that requires only $\Theta(\sqrt{n})$ samples.

3 Model

We will use the following model: Let $[n] := \{1, \ldots, n\}$. Then we will assume that $p$ is some black-box distribution over $[n]$, and that we can draw independent samples from $p$. We will also let $p_i := \text{Prob}[p \text{ outputs } i]$. We assume that we don’t know the $p_i$’s in advance.

Our goal will be to design an algorithm that either passes or fails a given probability distribution. It should pass the uniform distribution, and it should fail distributions that are far from uniform.

We measure uniformity in terms of the $l_1$ norm $\| \cdot \|_1$:

$$\|p - q\|_1 = \sum_i |p_i - q_i|$$

In particular, the distance of a distribution $p$ from the uniform distribution $U$ is

$$\|U - p\|_1 = \sum_i \left| \frac{1}{n} - p_i \right|$$

We will also have occasion to use other metrics. In general, we can consider the $l_p$ norm:
\[ \|p - q\|_p = \sqrt{\sum_i |p_i - q_i|^p} \]

In this lecture we will only look at the \( l_1 \) and \( l_2 \) norms.

The precise requirement we will ask of our algorithm is: The algorithm should pass the uniform distribution \( U \) and fail all distributions \( p \) such that \( \|p - U\|_1 > \epsilon \).

**Examples in \( l_1 \) and \( l_2 \)**

We will consider a few examples of measuring the distance from the uniform distribution under both the \( l_1 \) and \( l_2 \) norms. The purpose is to point out that your answer can change quite significantly based on the choice of norm.

**Example 1:** \( p_1 = 1, p_i = 0 \) for \( i \in \{2, \ldots, n\} \).

Then \( \|p - U\|_1 = (1 - \frac{1}{n}) + (n - 1) \frac{1}{n} = 2 \left(1 - \frac{1}{n}\right) \). Also, \( \|p - U\|_2 = \sqrt{(1 - \frac{1}{n})^2 + (n - 1) \frac{1}{n^2}} = \sqrt{1 - \frac{1}{n}} \). Thus in this case, \( \|p - U\|_1 \) and \( \|p - U\|_2 \) are within a factor of 2 of each other.

**Example 2:** \( p_i = \frac{2}{n} \) for \( i \in \{1, \ldots, \frac{n}{2}\} \), and \( p_i = 0 \) for \( i > \frac{n}{2} \).

Then \( \|p - U\|_1 = \frac{n}{2} \frac{1}{n} + \frac{n}{2} \frac{1}{n} = 2 \). But \( \|p - U\|_2 = \sqrt{\frac{n}{2} \frac{1}{n^2} + \frac{n}{2} \frac{1}{n^2}} = \frac{1}{n} \). Now the \( l_1 \) and \( l_2 \) norms differ by a factor of \( \frac{1}{n} \). In fact, \( p \) and \( U \) seem to be close in the \( l_2 \) norm, even though they are actually quite different.

This shows that just because \( \|p - U\|_2 \) is small doesn’t mean that we can infer that \( p \) is close to uniform. However, we have the following result:

\[ \|x\|_p \leq n^{\frac{1}{p} - \frac{1}{q}} \|x\|_q \]

In particular, with \( p = 1 \) and \( q = 2 \) we get that \( \|p - U\|_1 \leq \sqrt{n} \|p - U\|_2 \), so if \( \|p - U\|_2 << \frac{1}{\sqrt{n}} \) then \( \|p - U\|_1 << 1 \). Therefore it would suffice to come up with an algorithm that passes \( U \) and fails all distributions \( p \) such that \( \|p - U\|_2 > \frac{\epsilon}{\sqrt{n}} \).

### 4 Uniformity and \( \|p\|_2 \)

We next make a few observations that show a useful connection between uniformity and the \( l_2 \) norm.

First, we note that the distance from uniformity in the \( l_2 \) metric has a nice expression:

\[ \|p - U\|_2^2 = \sum_i \left(p_i - \frac{1}{n}\right)^2 = \sum_i p_i^2 - \frac{2}{n} \sum_i p_i + n \frac{1}{n^2} = \|p\|_2^2 - \frac{1}{n} \cdot \]

So, \( p \) is close to \( U \) in the \( l_2 \) norm if \( \|p\|_2^2 \) is close to \( \frac{1}{n} \).

Our second observation is that \( \|p\|_2^2 \) is equal to the probability that two samples from \( p \) end up being the same number, since the probability that both samples are \( i \) is \( p_i^2 \), and these events are all disjoint. We call \( \|p\|_2^2 \) the *collision probability*. We note again that \( \|p\|_2^2 \) can be computed as \( \sum_{i=1}^n p_i^2 \).

Also note that the uniform distribution is the unique distribution with the lowest possible collision probability \( -\frac{1}{n} \).
5 Strategy

We now get to our actual strategy. We will end up drawing $s$ samples from $p$, for some $s$, and looking at the number of pairs of these samples that are the same. We will use this to get a good estimate for $\|p\|_2^2$. Then, on the basis of that estimate, we will either accept or reject the distribution $p$ (based on whether it is sufficiently close to $\frac{1}{n}$).

A few comments: first, there are multiple senses in which we could get a good approximation to $\|p\|_2^2$ — either in an additive sense, or in a multiplicative sense (or even some other sense, but these are the two most common). Secondly, we only need to get a good enough approximation with probability greater than $\frac{1}{2}$, as once we do this, we can run our algorithm multiple times and take the majority vote on whether to accept or reject. This will make our success rate as close to 1 as we want it with only a logarithmic increase in number of samples.

Here is an overview of the algorithm:

1. Take $s$ samples from $p$.
2. Estimate $\|p\|_2^2$ (as the number of equal pairs in the sample, divided by $\binom{s}{2}$).
3. If $\|p\|_2^2 < \frac{1}{n} + \delta$, where $\delta$ is the error in the preceding estimate, then pass the distribution; otherwise, fail it.

This will never accidentally reject the uniform distribution, and as long as $\delta$ is sufficiently small, it will always reject distributions that are far from uniform. It turns out that we will need to choose $s = O\left(\frac{\epsilon}{\epsilon^2} \right)$ with $\delta = \frac{\epsilon^2}{2n}$.

We will go into more detail on how to actually implement and analyze this algorithm in the next lecture.

6 Additive and Multiplicative Estimates

We now go into a bit more detail on what sort of estimates we would need to construct our desired algorithm. There are two sorts of guarantees we will consider:

- **Guarantee 1**: If $p = U$ then pass, if $\|p - U\|_2 > \epsilon$ then fail.
- **Guarantee 2**: If $p = U$ then pass, if $\|p - U\|_1 > \epsilon$ then fail.

In both of these cases, we will consider what estimate $\hat{\|p\|}_2^2$ of $\|p\|_2^2$ we will need in order to achieve the desired guarantee.

In the case of Guarantee 1, recall from Section 4 that $\|p - U\|_2^2 = \|p\|_2^2 - \frac{1}{n}$. Therefore, if $\|p - U\|_2 > \epsilon$ then $\|p\|_2^2 > \frac{1}{n} + \epsilon^2$. On the other hand, $\|U\|_2^2 = \frac{1}{n}$. It follows that an additive estimate of $\|p\|_2^2$ to within $\frac{\epsilon^2}{2}$ will suffice to distinguish between the case of $p = U$ and $\|p - U\|_2 > \epsilon$.

In particular, we can adopt the following algorithm:

- Estimate $\|p\|_2^2$ to within $\frac{\epsilon^2}{2}$ additive error.
- If $\hat{\|p\|}_2^2 > \frac{1}{n} + \frac{\epsilon^2}{2}$, reject, otherwise, pass.
In the case of Guarantee 2, the story is a bit more complicated. Note that if \( \|p - U\|_1 > \epsilon \) then \( \|p - U\|_2 > \frac{\epsilon^2}{\sqrt{n}} \), so \( \|p - U\|_2^2 > \frac{\epsilon^2}{n} \). But this implies that \( \|p\|_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n} \). A multiplicative estimate for \( \|p\|_2^2 \) to within \( \frac{\epsilon^2}{3} \) would then imply that if \( \|p\|_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n} \) then our estimate \( \|p\|_2^2 \) will be at least

\[
\left( \frac{1}{n} + \frac{\epsilon^2}{n} \right) \left( 1 - \frac{\epsilon^2}{3} \right) = \frac{1 + \frac{2}{3} \epsilon^2 - \frac{1}{3} \epsilon^4}{n} > \frac{1 + \frac{1}{3} \epsilon^2}{n} = \frac{1}{n} \left( 1 + \frac{\epsilon^2}{3} \right)
\]

On the other hand, \( \|U\|_2^2 = \frac{1}{n} \), and so \( \|U\|_2^2 \) will be at most \( \frac{1}{n} \left( 1 + \frac{\epsilon^2}{3} \right) \). In particular, we can distinguish the uniform distribution from distributions with \( \|p - U\|_2^2 > \frac{\epsilon^2}{n} \), which means we can distinguish the uniform distribution from distributions with \( \|p - U\|_1 > \epsilon \). The particular algorithm is as follows:

- Estimate \( \|p\|_2^2 \) to within a multiplicative factor of \( 1 \pm \frac{\epsilon^2}{3} \).
- If \( \|p\|_2^2 > \frac{1}{n} \left( 1 + \frac{\epsilon^2}{3} \right) \), reject, otherwise, pass.

In summary, depending on what guarantee we want, we can search for either an additive or multiplicative bound on the error, and in both cases construct an algorithm that distinguishes uniform distributions from distributions that are far from uniform.