Lecture 11: Learning and testing probability distributions

In all we have done so far in this unit, we have assumed that the input was deterministic: the algorithm is randomised, yes, but the input itself is fixed, and arbitrary.

In this lecture, we (somewhat) change this. What we have is access to a sequence of independent, identically distributed (i.i.d.) data points, coming from an *unknown* probability distribution $\mathbf{p}$:

$$x_1, \ldots, x_n \sim \mathbf{p}$$

and what we want to do is to *learn something about this* $\mathbf{p}$. Put differently:

> The input is not the i.i.d. sequence $(x_1, \ldots, x_n)$: the input is $\mathbf{p}$, and $x_1, \ldots, x_n$ is how we get to access this input.

We will make very few assumptions about this unknown probability distribution $\mathbf{p}$: except that it is over a known *discrete* domain $\mathcal{X}$ of size $|\mathcal{X}| = k$.

To illustrate this, here is a histogram, corresponding to the counts from $n = 3665$ i.i.d. draws from some unknown probability distribution $\mathbf{p}$ over $\mathcal{X} = \{1, 2, \ldots, 49\}$ of size $k = 49$. They correspond to a number drawn, every week from 1982 to 2018, from Canada's "Lotto 6/49": Here are some questions we may want to

Presumably i.i.d.

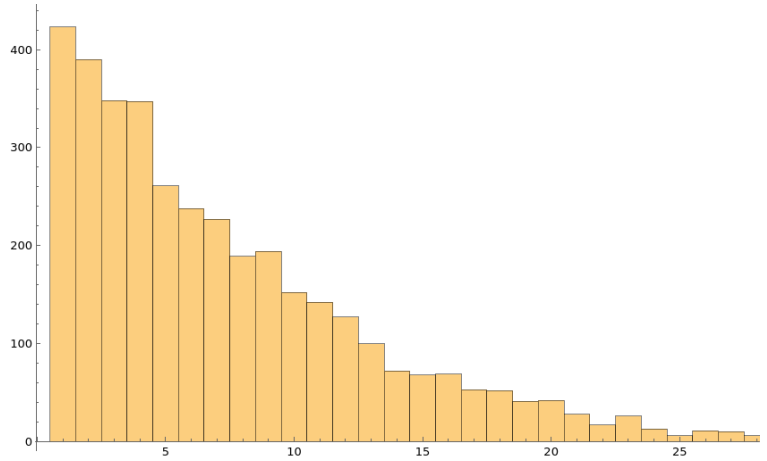

Figure 19: Counts for each of the $k = 49$ possible numbers among the $n = 3665$ draws. What distribution $\mathbf{p}$ does this correspond to? (Data from the Kaggle "Lotto 649" dataset.)

answer about $\mathbf{p}$:

- *What is it?* That is, can we *learn* the whole unknown probability distribution?
- *What are some of its characteristics?* That is, can we *learn* some "simple" parameter $f(\mathbf{p})$, such as its mean, entropy, variance?
- *Does it satisfy some specific property?* That is, can we *test* if $\mathbf{p}$ satisfies some requirement we care about? For instance, in the case of the lotto numbers above, "is $\mathbf{p}$ consistent with the distribution of the minimum of 6 independent uniform random numbers in $\{1, 2, \ldots, 49\}$?"

Intuitively, we can think of the first as learning $\approx k$ values about $\mathbf{p}$, the second as learning *one* value, and the last one as learning *one bit*. So presumably, they should be in decreasing order of "complexity," for whatever notion of "complexity" we define.

What is the notion of complexity? Our algorithms, as mentioned about, can only access the input $\mathbf{p}$ through queries which give *independent, identically distributed* samples from $\mathbf{p}$. We will of course care about the running time of the algorithms, but our main objective will be to minimise the number n of queries we make.

Think of n as the size of the dataset you need to collect, or generate, or buy in order for your algorithm to succeed. In the case of the lotto example, the number of observations is limited: there is only one lotto every week, we cannot choose n to be as large as we want.

What is the randomness? Since what we feed to the algorithm, the sequence of samples $x_1, \ldots, x_n$, is random, there will always be some probability our algorithm's output is wrong. That is, now, when we discuss expectations and probabilities, it will be over (1) the randomness of the algorithm itself, as usual, but also (2) the randomness in drawing $x_1, \ldots, x_n$ from the unknown $\mathbf{p}$.

What are the parameters? Of course, the domain size, k, is a key parameter of the problem. But we have at least two others: first, the probability of failure, δ: we want the algorithm to be correct about $\mathbf{p}$ most of the time. The other will be a *distance parameter*, $\varepsilon > 0$: we will get back to this soon, as its meaning depends on which of the three problems we are interested in. But overall, our goal will be:

> Find the smallest value $n = n(k, \varepsilon, \delta)$ for which an algorithm can solve the learning, estimation, or testing task we care about with distance parameter ε and failure probability at most δ.

Some notation. A probability distribution over $\mathcal{X}$ can be identified to its probability mass function (pmf), which is a function $\mathbf{p} : \mathcal{X} \to [0, 1]$ such that $\sum_{x \in \mathcal{X}} \mathbf{p}(x) = 1$. Accordingly, the probability mass that $\mathbf{p}$ assigns to a subset[45] $S \subseteq \mathcal{X}$ is $\mathbf{p}(S) = \sum_{x \in S} \mathbf{p}(x) = \Pr_{x \sim \mathbf{p}}[\, x \in S \,]$. Finally, $\Delta(\mathcal{X})$ denotes the set of all probability distributions over our domain $\mathcal{X}$.

[45] As we only consider discrete domains, we ignore issues of measurability, etc. That is, we consider $\mathcal{X}$ endowed with the counting measure, so every set is measurable. What is discussed here generalises to continuous probability distributions, but with annoying technical details.

Distance between probability distributions

To define what it means to learn a probability distribution, or even how *far* two probability distributions over the same domain $\mathcal{X}$ are, we need a notion of *distance*. Ideally, one which makes "sense": (1) a metric would be nice (to be able to use the triangle inequality when needed), (2) a *bounded* metric would be even nicer (to be able to understand a value such as 0.1 without having to normalise or think twice), (3) a bounded metric *with a simple and meaningful interpretation* would be best.

This leads us to the the notion of distance we will be concerned about, the total variation distance (also known as *statistical distance*).

Definition 49.1 (Total variation distance)**.** The *total variation distance* between two probability distributions $\mathbf{p}, \mathbf{q} \in \Delta(\mathcal{X})$ is given by

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \mathcal{X}} (\mathbf{p}(S) - \mathbf{q}(S)) .$$

Given a subset $\mathcal{C} \subseteq \Delta(\mathcal{X})$ of distributions, we further define the distance from $\mathbf{p} \in \Delta(\mathcal{X})$ to $\mathcal{C}$ as $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}) := \inf_{\mathbf{q} \in \mathcal{C}} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$, and will say that $\mathbf{p}$ is ε-*far from* $\mathcal{C}$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}) > \varepsilon$.

One can check that d_{TV} defines a metric on $\Delta(\mathcal{X})$, and takes values in $[0, 1]$. Moreover, the total variation distance exhibits several important properties, among which one of the most important is its *immunity to post-processing*:

Check it!

Fact 49.1 (Data Processing Inequality)**.** *Suppose X and Y are independent random variables with distributions $\mathbf{p}$ and $\mathbf{q}$, and let f be any (possibly randomized) function independent of X, Y. Then the probability distributions $\mathbf{p}_f$ and $\mathbf{q}_f$ of $f(X)$ and $f(Y)$ satisfy*

$$\mathrm{d}_{\mathrm{TV}}\left(\mathbf{p}_f, \mathbf{q}_f\right) \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) .$$

That is, postprocessing cannot increase the total variation distance.

What this says is, paraphrasing, that post-processing two random variables the same way cannot "make them statistically farther."

Interestingly, total variation distance also has a very natural interpretation in terms of *indistinguishability*:

Lemma 49.1 (Pearson–Neyman)**.** *Any (possibly randomized) algorithm which distinguishes between $\mathbf{p}$ and $\mathbf{q}$ from a* single *sample must have Type I (false positive) and Type-II (false negative) errors satisfying*

$$\textit{Type I} + \textit{Type II} \geq 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$$

Moreover, this is achieved by the test which outputs "$\mathbf{q}$" if, and only if, the sample belongs to the "Scheffé set" $S^* := \{\, x : \mathbf{q}(x) > \mathbf{p}(x) \,\}$.

You can ignore the proof in a first read, it is just here for completeness. What matters is the lemma itself.

Proof. Fix any test A distinguishing between two distributions $\mathbf{p}$ and $\mathbf{q}$, given a single observation. Letting α and β denote the Type I and Type-II errors of A, we have

$$\begin{aligned}\alpha + \beta &= \Pr_{\mathbf{p},R}[\, A(X,R) = 1\,] + \Pr_{\mathbf{q},R}[\, A(X,R) = 0\,] \\ &= \mathbb{E}_R[\Pr_{\mathbf{p}}[\, A(X,R) = 1\,]] + \mathbb{E}_R[\Pr_{\mathbf{q}}[\, A(X,R) = 0\,]] \\ &= \mathbb{E}_R[\Pr_{\mathbf{p}}[\, A(X,R) = 1\,] + \Pr_{\mathbf{q}}[\, A(X,R) = 0\,]]\end{aligned}$$

where we denote by R the internal randomness of A. Since, for any fixed realization r of this randomness R, the resulting test $A(\cdot, r)$ is deterministic, we can define for any r the *acceptance region* $S_{A,r} := \{\, x : A(x,r) = 1 \,\}$, and write

$$\begin{aligned}\alpha + \beta &= \mathbb{E}_R[\Pr_{\mathbf{p}}[\, X \in S_{A,R}\,] + \Pr_{\mathbf{q}}[\, X \notin S_{A,R}\,]] \\ &= 1 + \mathbb{E}_R[\mathbf{p}(S_{A,R}) - \mathbf{q}(S_{A,R})] \\ &\geq 1 + \inf_S(\mathbf{p}(S) - \mathbf{q}(S)) \\ &= 1 - \sup_S(\mathbf{q}(S) - \mathbf{p}(S)) \\ &= 1 - \mathrm{d_{TV}}(\mathbf{p},\mathbf{q})\,,\end{aligned}$$

as claimed. Finally, it is immediate from the definition of total variation distance that the proposed test satisfies Type I + Type II $= 1 + \mathbf{p}(S^*) - \mathbf{q}(S^*) = 1 - \mathrm{d_{TV}}(\mathbf{p},\mathbf{q})$. □

Here is one way to interpret this lemma:

> Alice and Bob play a game, where they both know two probability distributions $\mathbf{p}, \mathbf{q}$. Alice starts by tossing a fair coin, and does not show the outcome to Bob: if it is Heads, then she draws $x \sim \mathbf{p}$; if it is Tails, she draws $x \sim \mathbf{q}$. Then she shows the value of x to Bob, who must guess if the coin toss was Heads. Clearly, just by random guessing, Bob can win the game with probability $1/2$. What the lemma says is that he can do better: there is a strategy for him to win with probability
>
> $$\Pr[\,\text{Bob wins}\,] = \frac{1}{2} + \frac{\mathrm{d_{TV}}(\mathbf{p},\mathbf{q})}{2}$$
>
> and, moreover, this is the best possible.

One more (very useful) fact about total variation distance: it is just ℓ_1 distance in disguise!

Fact 49.2 (Scheffé's Lemma)**.** *For any two* $\mathbf{p}, \mathbf{q} \in \Delta(k)$*,*

$$\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathbf{p}(x) - \mathbf{q}(x)| = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1 \tag{68}$$

that is, "total variation is half the ℓ_1 distance between pmfs."

This fact, which you will prove during the tutorial, turns out to be a very useful connection: if nothing else, ℓ_p norms are well studied, and this will allow us to use our arsenal of geometric inequalities — Hölder, Cauchy–Schwarz, and monotonicity of ℓ_p norms, to name a few.

The case of a coin ($k = 2$)

With the necessary background in hand, we can look at our first question: forget for now about $k \gg 1$, let us focus on the simplest, most basic case, $k = 2$: you are given *a coin*, and it may be biased.

The learning task can be then rephrased as follows:

> How many times n do you need to flip the coin to learn its true bias p to accuracy $\pm\varepsilon$, and be correct with probability at least $1 - \delta$?

Should it be...

- $n = O\left(\frac{1}{\varepsilon\delta}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon^2\delta}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ times?

Well, actually, we can get something more refined than any of the bounds above! If we are given a promise on the unknown bias p, then we can get the following:

Theorem 50. *Suppose we are promised that the true bias p of the coin satisfies $0 \le p < q \le \frac{1}{2}$, for some known value q. Then estimating the bias of the coin to an additive ε, with probability at least $1 - \delta$, can be done with $n = O\left(\frac{q}{\varepsilon^2}\log\frac{1}{\delta}\right)$ i.i.d. samples. (Moreover, this is optimal.)*

Proof. This follows from a Chernoff bound, using the empirical estimate

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, \dots, x_n \sim \mathrm{Bern}(p)$. Note that a Hoeffding bound would not give you the dependence on q. (The lower bound, *i.e.*, proof of optimality, is beyond the scope of this lecture.) □

As a remark, the assumption that $q \in (0, 1/2]$ is without loss of generality: if instead we were promised that $q < p \le 1$ with $q \in (1/2, 1)$, then we could flip the coin flips (!) by looking at $x_i' := 1 - x_i$ instead, and estimate the bias $p' := 1 - p$, which satisfies $0 \le p' < 1 - q \le 1/2$. Clearly, estimating p' to $\pm\varepsilon$ is equivalent to estimating p to $\pm\varepsilon$.

Corollary 50.1. *Estimating the bias of a coin to an additive ε, with probability at least $1-\delta$, can be done with $n = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ i.i.d. samples. (Moreover, this is optimal.)*

This follows from applying the above theorem setting $q = 1/2$; alternatively, the upper bound can be directly proven using the Hoeffding bound.

Try it!

This is if we want to *learn* the bias of an unknown coin. What if we just want to *test* if it is biased? That is, distinguish between the case where (1) the coin is Heads with probability exactly $1/2$ (fair coin), or probability $1/2 \pm \Omega(\varepsilon)$ (biased coin)? How many times n would we have to toss the coin, if we wanted our diagnostic to be correct with probability at least $1-\delta$?

- $n = O\left(\frac{1}{\varepsilon}\sqrt{\log\frac{1}{\delta}}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon^2}\sqrt{\log\frac{1}{\delta}}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ times?
- $n = O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ times?

As it turns out... *testing* whether the coin is biased or fair is basically as hard as *learning* the bias of the coin:

Theorem 51. *Testing whether the bias of a coin is $1/2$ or at least $1/2+\varepsilon$, with probability at least $1-\delta$, can be done with $n = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ i.i.d. samples. (Moreover, this is optimal.)*

This is a little underwhelming, since one can literally do this by learning the bias up to $\frac{\varepsilon}{2}$.[46] And that provides a lot more information! So is there anything to be gained (except maybe constant factors) if we only want to test the bias of the coin? As it turns out, *yes*... but not always. *Not* when testing if a coin is fair or biased: but when testing if the coin is very biased or *extremely* biased, then yes.

[46] Can you see how?

Theorem 52. *For any $0 < \alpha \le 1/2$ and $\varepsilon \in (0,1]$, testing whether the bias of a coin is at most α or at least $\alpha(1+\varepsilon)$, with probability at least $1-\delta$, can be done with $n = O\left(\frac{1}{\alpha\varepsilon^2}\log\frac{1}{\delta}\right)$ i.i.d. samples.*

Again, this is not very useful when $\alpha = \Omega(1)$: however, for vanishing α, this is much better than what learning the bias to an additive $\pm\frac{1}{2}\alpha\varepsilon$ would give, which by Corollary 50.1 is $n = O\left(\frac{1}{\alpha^2\varepsilon^2}\log\frac{1}{\delta}\right)$.

Theorem 52 can be proven by a Chernoff bound, specifically the version given in Theorem 14 applied to the same empirical estimator $\hat{p}$. But rather than *proving* it, let us give a small sketch of *why* we could expect this statement to be true, to get some intuition (and remember the result). If the true bias p is roughly $\Theta(\alpha)$, then we expect to see Tails most of the time, and Heads a $\Theta(\alpha)$ fraction of the tosses. That is, in every chunk of $1/2\alpha$ tosses, we expect to

see a Heads with probability either at most $1/2$ (if $p \leq \alpha$) or at least $(1+\varepsilon)/2$ (if $p \geq (1+\varepsilon)\alpha$). But by Theorem 51 (ignoring δ for simplicity), this takes $O(1/\varepsilon^2)$ chunks – so $O(1/(\alpha\varepsilon^2))$ coin tosses in total.

And this makes sense! Things should become easier in the "highly biased" regime. With a similar analysis, one can show that distinguishing between bias $p = 0$ and bias $p \geq \varepsilon$ takes only $n = O((1/\varepsilon)\log(1/\delta))$ coin tosses. And this is easy to interpret: in one case, you *never* see a Heads, and in the other, you will see one after $\approx 1/\varepsilon$ coin tosses. What is really hard (and requires more coin tosses) is when $p \approx 1/2$, and you have to distinguish between "a lot of Heads" and "a lot of Heads, *but slightly more*."

Learning and testing beyond coins

This is all well and good, but we often have to consider data over domains of size $k \gg 2$. The lotto example above, for instance, was for $k = 49$; and that's only for *one* number: if one considers all 6 draws in a single ticket of that Canadian lotto, that's a domain of size $k = 49^6 = 13,841,287,201$.

If we were given an algorithm to generate random permutations and we wanted to test whether its output was truly uniform (on the space of all permutations of, say, size 8), then we would be looking at a space of size $16! = 209,22,789,888,000$.

If we wanted to estimate the entropy of a dataset of 8-character passwords made of lower and uppercase letters, digits, and special characters \$%#&!?_−, then $k = (70)^8 = 576,480,100,000,000$.

> Domain sizes grow quite fast, and in most settings k is huge.

If we cannot assume structure in the data, then we have to hope for *very* sample-efficient algorithms.

Learning

In *learning* (in total variation distance),[47] our goal is to design an algorithm A which, given n i.i.d. samples from $\mathbf{p}$ and parameters $\varepsilon, \delta \in (0,1]$, outputs $\widehat{\mathbf{p}}$ such that

$$\Pr[\,\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) > \varepsilon\,] \leq \delta \tag{69}$$

that is, $\widehat{\mathbf{p}}$ is close to $\mathbf{p}$, with high probability. The probability is taken, again, over both the samples $x_1, \ldots, x_n \sim \mathbf{p}$ *and* the randomness of the algorithm A.

How many samples n would we have to take, if we wanted the output $\widehat{\mathbf{p}}$ to be ε-close to the true $\mathbf{p}$ with probability at least $1-\delta$?

- $n = O\left(\frac{k^2}{\varepsilon}\log\frac{1}{\delta}\right)$?
- $n = O\left(\frac{k}{\varepsilon^2}\log\frac{1}{\delta}\right)$?

[47] One can define learning with respect to other notions of distances: *e.g.*, Kullback–Leibler divergence, or ℓ_∞ distance. Here we focus on the standard, nice, total variation.

- $n = O\left(\frac{k+\log\frac{1}{\delta}}{\varepsilon^2}\right)$?
- $n = O\left(\frac{k^2+\log\frac{1}{\delta}}{\varepsilon^2}\right)$?

As we will see, the answer is not entirely obvious, even though the algorithm A is.

First idea: using what we saw. We want to estimate all k probabilities $\mathbf{p}_1, \dots, \mathbf{p}_k$ to get $\widehat{\mathbf{p}}_1, \dots, \widehat{\mathbf{p}}_k$ such that

$$\frac{1}{2}\sum_{i=1}^{k} |\widehat{\mathbf{p}}_i - \mathbf{p}_i| \le \varepsilon\,.$$

It would be *enough* to estimate each individual $\mathbf{p}_i$ to an additive $\frac{2\varepsilon}{k}$.[48] To make sure we learn *all* k of them, we will learn each with error failure $\frac{\delta}{k}$ and take a union bound (using the same set of n samples for all k estimates).

[48] In case you are worried we may not end up with $\sum_{i=1}^{k} \widehat{\mathbf{p}}_i = 1$: since we are using the same n samples for all k estimates, we get "for free" that the sum is 1.

The total cost, from Corollary 50.1, is then

$$n = O\left(\frac{1}{(\varepsilon/k)^2}\log\frac{1}{(\delta/k)}\right) = O\left(\frac{k^2}{\varepsilon^2}\log\frac{k}{\delta}\right). \tag{70}$$

That's something, but that has a more-than-*quadratic* dependence on this giant parameter k.

But we can do better! Another idea would be to learn each $\mathbf{p}_i$ to a multiplicative factor $(1 \pm 2\varepsilon)$, instead of an additive $\pm\frac{2\varepsilon}{k}$. *If* we assume that $\mathbf{p}_i \ge \frac{\varepsilon}{k}$ for all $1 \le i \le k$, for instance, then a Chernoff bound (along with a union bound) tell us that the empirical estimates $\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_k$ satisfy

$$(1-2\varepsilon)\mathbf{p}_i \le \widehat{\mathbf{p}}_i \le (1+2\varepsilon)\mathbf{p}_i, \text{ for all } 1 \le i \le k$$

with probability at least $1-\delta$, for

$$n = O\left(\frac{k}{\varepsilon^3}\log\frac{k}{\delta}\right) \tag{71}$$

and then

$$\frac{1}{2}\sum_{i=1}^{k} |\widehat{\mathbf{p}}_i - \mathbf{p}_i| \le \frac{1}{2}\sum_{i=1}^{k} 2\varepsilon\mathbf{p}_i = \varepsilon\,,$$

since $\sum_{i=1}^{k} \mathbf{p}_i = 1$. Moreover, we can get rid of that assumption on $\min_i \mathbf{p}_i$, losing only constant factors in the final bound.

($\star$) Can you see how? Hint: "mix" $\mathbf{p}$ with uniform, and learn

$$\mathbf{p}' = (1-\frac{\varepsilon}{2})\mathbf{p} + \frac{\varepsilon}{2}\mathbf{u}_k$$

instead, where $\mathbf{u}_k$ is the uniform distribution on $\mathcal{X}$.

But we can do better! One of the two bounds above has a (near) quadratic dependence on k but a quadratic dependence on $1/\varepsilon$, the other is (near) linear in k but has a cubic dependence on $1/\varepsilon$, and both have an extra logarithmic factor in k because of a union bound. This does not "feel" right, and indeed it is not:

Theorem 53. *Learning an unknown distribution $\mathbf{p} \in \Delta(k)$ to total variation distance ε (with success probability $1-\delta$) can be done with*

$$n = O\left(\frac{k+\log\frac{1}{\delta}}{\varepsilon^2}\right)$$

i.i.d. samples. (Moreover, this is optimal.)

We will only prove the upper bound statement (not the lower bound showing this sample complexity is optimal), but it is worth noting that we have $k + \log(1/\delta)$, *not* $k \log(1/\delta)$: this is perhaps surprising, as $k \log(1/\delta)$ is what the median trick would give us.

Proof of Theorem 53. Consider the empirical distribution $\widehat{\mathbf{p}}$ obtained by drawing n independent samples $x_1, \dots, x_n$ from the underlying distribution $\mathbf{p} \in \Delta([k])$:

$$\widehat{\mathbf{p}}(i) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{x_j = i\}}, \qquad i \in [k] \tag{72}$$

This defines a valid probability distribution (*i.e.*, $\widehat{\mathbf{p}} \in \Delta(k)$), and moreover can be computed efficiently, in time $O(k + n \log n)$.

Recalling the definition of total variation distance (Definition 49.1), the key observation is that we have $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) > \varepsilon$ if, and only if, there exists a subset $S \subseteq [k]$ such that $\widehat{\mathbf{p}}(S) > \mathbf{p}(S) + \varepsilon$. There are only 2^k subsets (actually $2^k - 2$) to consider, so we will make sure our estimate is accurate for *each* subset, taking a union bound over all 2^k of them.

Fix any $S \subseteq [k]$. We have

$$\widehat{\mathbf{p}}(S) = \sum_{i \in S} \widehat{\mathbf{p}}(i) \overset{(72)}{=} \frac{1}{n} \sum_{i \in S} \sum_{j=1}^{n} \mathbb{1}_{\{x_j = i\}}$$

and so, letting $X_j := \sum_{i \in S} \mathbb{1}_{\{x_j = i\}}$ for $j \in [n]$, we have $\widehat{\mathbf{p}}(S) = \frac{1}{n} \sum_{j=1}^{n} X_j$ where the X_j's are i.i.d. Bernoulli random variables with parameter $\mathbf{p}(S)$. By a Hoeffding bound,

$$\Pr[\, \widehat{\mathbf{p}}(S) > \mathbf{p}(S) + \varepsilon \,] = \Pr\left[\frac{1}{n} \sum_{j=1}^{n} X_j > \mathbb{E}\left[\frac{1}{n} \sum_{j=1}^{n} X_j \right] + \varepsilon \right] \leq e^{-2\varepsilon^2 n}$$

and therefore $\Pr[\, \widehat{\mathbf{p}}(S) > \mathbf{p}(S) + \varepsilon \,] \leq \frac{\delta}{2^k}$ as long as

$$n \geq \frac{k \ln 2 + \log(1/\delta)}{2\varepsilon^2} \tag{73}$$

A union bound over these 2^k possible sets S concludes the proof:

$$\Pr[\, \exists S \subseteq [k] \text{ s.t. } \widehat{\mathbf{p}}(S) > \mathbf{p}(S) + \varepsilon \,] \leq 2^k \cdot \frac{\delta}{2^k} = \delta\,. \qquad \square$$

This proof is a little magical, and crucially relies on the definition of total variation distance as a supremum over subsets. One can also prove the statement using the equivalent characterisation (from Fact 49.2) as ℓ_1 distance. We will only prove it for constant δ, as the full version requires a tool (McDiarmid's inequality) we have not seen in this class.

Alternative proof of Theorem 53. Consider the empirical distribution $\widehat{\mathbf{p}}$ from n i.i.d. samples defined in (72).

First, we bound the *expected* total variation distance between $\widehat{\mathbf{p}}$ and $\mathbf{p}$, by using ℓ_2 distance as a proxy:

$$\mathbb{E}[\mathrm{d_{TV}}(\mathbf{p},\widehat{\mathbf{p}})] = \frac{1}{2}\mathbb{E}[\|\mathbf{p}-\widehat{\mathbf{p}}\|_1] = \frac{1}{2}\sum_{i=1}^{k}\mathbb{E}[|\mathbf{p}(i)-\widehat{\mathbf{p}}(i)|]$$
$$\leq \frac{1}{2}\sum_{i=1}^{k}\sqrt{\mathbb{E}[(\mathbf{p}(i)-\widehat{\mathbf{p}}(i))^2]}$$

the last inequality by Jensen. But since, for every $i \in [k]$, $n\widehat{\mathbf{p}}(i)$ follows a $\mathrm{Bin}(n,\mathbf{p}(i))$ distribution, we have

$$\mathbb{E}\left[(\mathbf{p}(i)-\widehat{\mathbf{p}}(i))^2\right] = \frac{1}{n^2}\mathrm{Var}[n\widehat{\mathbf{p}}(i)] = \frac{1}{n}\mathbf{p}(i)(1-\mathbf{p}(i))$$

from which

$$\mathbb{E}[\mathrm{d_{TV}}(\mathbf{p},\widehat{\mathbf{p}})] \leq \frac{1}{2\sqrt{n}}\sum_{i=1}^{k}\sqrt{\mathbf{p}(i)} \leq \frac{1}{2}\sqrt{\frac{k}{n}}$$

the last inequality this time by Cauchy–Schwarz. Therefore, for $n \geq \frac{25k}{\varepsilon^2}$ we have $\mathbb{E}[\mathrm{d_{TV}}(\mathbf{p},\widehat{\mathbf{p}})] \leq \frac{\varepsilon}{10}$.

We can then conclude by Markov's inequality, establishing that

$$\Pr[\,\mathrm{d_{TV}}(\mathbf{p},\widehat{\mathbf{p}}) \geq \varepsilon\,] \leq \frac{1}{10}\,. \qquad \square$$

($\star$) A final (and side) remark: this last proof actually establishes a slightly stronger result, namely, that the sample complexity n can be expressed as $n = O(\|\mathbf{p}\|_{1/2}/\varepsilon^2)$, where $\|\mathbf{p}\|_{1/2} = \left(\sum_{i=1}^{k}\sqrt{\mathbf{p}(i)}\right)^2$ is the "$\frac{1}{2}$-norm" of the unknown distribution $\mathbf{p}$.

This is (1) not examinable, and (2) not a norm.

Testing

As we just saw, we can learn the *whole* probability distribution $\mathbf{p}$ to accuracy ε using $O(k/\varepsilon^2)$ samples. What if we only wanted to test if $\mathbf{p}$ had some important property? For instance, if $\mathbf{p} = \mathbf{q}$, where $\mathbf{q} \in \Delta(k)$ is some known reference distribution?

Specifically, we want to solve the following problem:

This question is called "identity testing" in the distribution testing literature.

> Give an algorithm A which takes parameters $\varepsilon, \delta \in (0,1]$ and n samples from $\mathbf{p}$, and:
>
> - If $\mathbf{p} = \mathbf{q}$, then $\Pr[\,A$ outputs yes$\,] \geq 1-\delta$;
> - If $\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) > \varepsilon$, then $\Pr[\,A$ outputs no$\,] \geq 1-\delta$
>
> (if $0 < \mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) \leq \varepsilon$, then A is off the hook and can output whatever).

Why do we care? For instance, someone may hand you an algorithm claiming it samples a uniformly random permutation; or the implementation of a hash family $\mathcal{H}$, claiming its output $h(x)$ (over the choice of $h \sim \mathcal{H}$) is uniformly distributed for each x; or you

may have an algorithm running really well on uniformly random data, but very poorly on very skewed data – and you want to test these claims, or if your dataset is uniform enough for your fast algorithm.

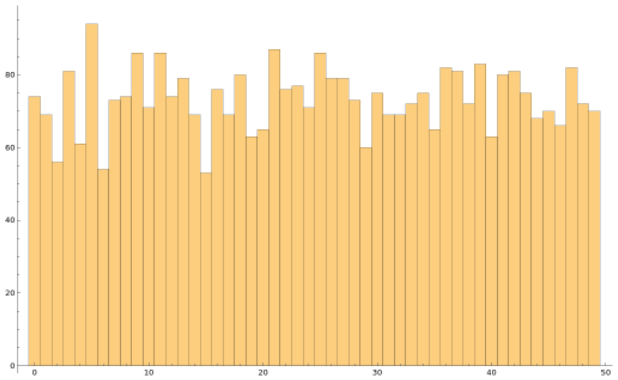

Figure 20: Histogram of 3,665 draws of the "bonus number" in Canada's 6/49 lotto, each draw being a number in $\{1, 2, \ldots, 49\}$. Is the distribution uniform? Or is the lottery not fair?

So you have a reference distribution $\mathbf{q} \in \Delta(k)$ in mind. The first thing we will do is simplify the problem, and assume that $\mathbf{q}$ is not *any* distribution, but *the* uniform distribution $\mathbf{u}_k$ over $\mathcal{X}$. This will make our life easier. And while this seems like a big simplification, it turns out it is not! That is, any algorithm for *uniformity testing* (the reference is $\mathbf{u}_k$) can be used as a blackbox to solve the *identity testing* (the reference is any $\mathbf{q}$), *via* a reduction:

This is the *uniformity testing* question, a special case of identity testing.

Theorem 54 (Identity to uniformity reduction)**.** *Suppose there is an algorithm A for uniformity testing, which takes $n = n(k, \varepsilon, \delta)$ i.i.d. samples from the unknown distribution. Then there is an algorithm A' for identity testing over a domain of size k to any fixed $\mathbf{q} \in \Delta(k)$, which takes $n = n(4k, \varepsilon/4, \delta)$ i.i.d. samples from the unknown distribution. Moreover, A' is efficient if A is.*

We will not prove this theorem here, but this essentially says that while identity testing seems to be a strict generalization of uniformity testing, in terms of number of samples required they are basically equivalent (up to constant factors).

Now, all we need to solve is the *uniformity testing* problem: given n i.i.d. samples from an unknown $\mathbf{p}$ over $\mathcal{X}$, decide whether $\mathbf{p} = \mathbf{u}_k$, or $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$ (and be correct with probability at least $1 - \delta$). Let's say $\delta = 1/3$. How many samples n would we have to take to solve this question?

- $n = O\left(\frac{k}{\varepsilon^2}\right)$?
- $n = O\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$?
- $n = O\left(\frac{1}{\varepsilon^2}\right)$?
- $n = O\left(\frac{\log k}{\varepsilon^2}\right)$?

In what follows, we will assume $\delta = 1/3$, since we can boost this to any $1 - \delta$ by a "standard majority vote" losing only an $O(\log(1/\delta))$ factor in the number of samples n.

Exercise: write down the details!

A baseline: if we can learn, we can test. The first claim is that the sample complexity of *learning* is an upper bound on that of *testing*: that is, one can always to do the following.

- Learn $\mathbf{p}$ to total variation distance $\frac{\varepsilon}{2}$ to obtain $\widehat{\mathbf{p}}$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) \leq \frac{\varepsilon}{2}$ with probability at least $2/3$;
- Check (without taking any more samples) if $\mathrm{d}_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{u}_k) \leq \frac{\varepsilon}{2}$;
- Output `yes` if it is the case, `no` otherwise.

Since total variation distance is a metric, it satisfies the triangle inequality: so

- If $\mathbf{p} = \mathbf{u}_k$, then $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) \leq \frac{\varepsilon}{2}$ (with probability at least $2/3$);
- If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, then by the triangle inequality

$$\varepsilon < \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) + \mathrm{d}_{\mathrm{TV}}(\widehat{\mathbf{p}}, \mathbf{u}_k)$$

  but by our learning guarantee the first term is at most $\varepsilon/2$ (with probability at least $2/3$), so the second must be more than $\varepsilon/2$.

This simple argument tells us that whatever we end up getting, we should do no worse than $n = O(k/\varepsilon^2)$: since that is what the learning approach will get us.

But we can do better! Alright, we can do $n = O(k)$ by learning: but again, here we only aim for *one bit* of information. As it turns out, this allows us to do significantly better in terms of sample complexity:

Theorem 55. *Testing uniformity of an unknown distribution $\mathbf{p} \in \Delta(k)$ to total variation distance ε (with success probability $2/3$) can be done with*

$$n = O\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$$

i.i.d. samples, using Algorithm 23. (Moreover, this is optimal for constant success probability.)

In terms of dependence on k, this is a *quadratic* improvement over learning! Before giving (part of) the proof, you may wonder where this $\sqrt{k}$ comes from: at a high level, it comes from something we have seen before, the *Birthday Paradox*. Consider the distribution $\mathbf{p}$ which is uniform on an arbitrary subset of $k/2$ elements: it is easy to see that it is at total variation distance $1/2$ from $\mathbf{u}_k$. But unless we take $n = \Omega(\sqrt{k})$ samples from $\mathbf{p}$, all we see is a sequence on unique elements from the domain, with zero collisions: which is entirely, and absolutely consistent with what we would see under the uniform distribution, too!

Why $\sqrt{k}$? Birthday Paradox.

We will only show here how to derive a (suboptimal) bound $n = O\left(\sqrt{k}/\varepsilon^4\right)$.

(Partial) proof of Theorem 55. This idea that *collisions* are important to test whether $\mathbf{p}$ is uniform is actually quite important, and the basis behind Algorithm 23. Namely, we will use the following facts:

Algorithm 23: COLLISION-BASED UNIFORMITY TESTER

Input: Multiset of n i.i.d. samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0,1]$ and $k = |\mathcal{X}|$

Set $\tau \leftarrow \frac{1+2\varepsilon^2}{k}$

Compute ▷ $O(n)$ time if $\mathcal{X}$ is known

$$Z = \frac{1}{\binom{n}{2}} \sum_{1 \le s < t \le n} \mathbb{1}_{\{x_s = x_t\}} = \frac{1}{\binom{n}{2}} \sum_{j \in \mathcal{X}} \binom{N_j}{2}$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}_{\{x_t = j\}}$.

if $Z \ge \tau$ **then return** no ▷ Not uniform

else return yes ▷ Uniform

the first is what while TV distance is basically ℓ_1 distance between pmfs, the ℓ_2 *distance is a good proxy for total variation distance*:

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) = \frac{1}{2}\|\mathbf{p} - \mathbf{u}_k\|_1 \le \frac{\sqrt{k}}{2}\|\mathbf{p} - \mathbf{u}_k\|_2 \tag{74}$$

the inequality being Cauchy–Schwarz. What this means is that

- if $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 > 4\varepsilon^2/k$; while
- if $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) = 0$ then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 = 0$ too.

So it is *sufficient* to test with respect to ℓ_2 distance. What does that buy us? We have the very convenient fact, specific to the distance to the uniform distribution: for any distribution $\mathbf{p}$ over $\mathcal{X}$,

$$\|\mathbf{p} - \mathbf{u}_k\|_2^2 = \sum_{i=1}^{k} \left(\mathbf{p}(i) - \frac{1}{k}\right)^2 = \sum_{i=1}^{k} \mathbf{p}(i)^2 - \frac{1}{k} = \|\mathbf{p}\|_2^2 - \frac{1}{k}, \tag{75}$$

so combining the two we get that $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$ implies $\|\mathbf{p}\|_2^2 > (1 + 4\varepsilon^2)/k$.

Remark 55.1. We have seen this quantity $\|\mathbf{p}\|_2^2$ before! It is commonly known as the *collision probability* of $\mathbf{p}$, due to the following fact: if X, Y are i.i.d. random variables distributed according to $\mathbf{p}$, then

$$\Pr[X = Y] = \sum_{i \in \mathcal{X}} \Pr[X = i, Y = i] = \sum_{i \in \mathcal{X}} \mathbf{p}(i)^2 = \|\mathbf{p}\|_2^2 \tag{76}$$

It is easy to see, from Eq. (75), that among all probability distributions over a given support size k the collision probability is minimised for the uniform distribution: indeed, $\|\mathbf{p}\|_2^2 = \frac{1}{k} + \|\mathbf{p} - \mathbf{u}_k\|_2^2 \ge \frac{1}{k}$.

In view of Eq. (75), a very natural idea is to estimate $\|\mathbf{p}\|_2^2$, in order to distinguish between (i) $\|\mathbf{p}\|_2^2 = 1/k$ (uniform) and (ii) $\|\mathbf{p}\|_2^2 > (1 + 4\varepsilon^2)/k$ (ε-far from uniform). How to do that? We just saw that the probability that two independent samples from $\mathbf{p}$ take the same value (a "collision") is exactly $\|\mathbf{p}\|_2^2$. Thus, an obvious approach is to take n samples $x_1, \dots, x_n$, count the number of pairs that show a collision, and use that as an estimator Z for $\|\mathbf{p}\|_2^2$:

$$Z = \frac{1}{\binom{n}{2}} \sum_{1 \le s < t \le n} \mathbb{1}_{\{x_s = x_t\}}. \tag{77}$$

Sanity check: why not just look at $n/2$ (independent) pairs of samples, and use them to estimate $\Pr[X = Y]$?

By the above, $\mathbb{E}[Z] = \|\mathbf{p}\|_2^2$. If we threshold Z somewhere between (i) and (ii), at say

$$\tau := \frac{1 + 2\varepsilon^2}{k}$$

we should be able to distinguish between our two cases and get a valid tester. But how large must n be for this to work?

Intuitively, we expect the test to work as long as the standard deviation of Z (the "noise") is smaller than the gap between the expectations in our two cases (the "signal"); that is,

$$\sqrt{\operatorname{Var}[Z]} \ll \Delta\mathbb{E}[Z] = \frac{4\varepsilon^2}{k} \tag{78}$$

as this is the condition for the random fluctuations of our statistic Z not to "cross" our threshold too often and lead to a wrong answer.

To make this quantitative, we can use Chebyshev's inequality, which requires us to bound $\operatorname{Var}[Z]$. This is where things get tricky, since Z is the sum of $\binom{n}{2}$ random variables which are *not* pairwise independent.[49]

We will only show here to derive a (suboptimal) bound $n = O\left(\sqrt{k}/\varepsilon^4\right)$:

$$\begin{aligned}\operatorname{Var}[Z] &= \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2 \\ &= \frac{1}{\binom{n}{2}^2} \sum_{1\le s<t\le n} \sum_{1\le s'<t'\le n} \mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\mathbb{1}_{\{X_{s'}=X_{t'}\}}\right] - \|\mathbf{p}\|_2^4\end{aligned}$$

To handle this last sum despite the lack of independence of the summands, we will break it in 3 groups depending on the cardinality of $\{s,t,s',t'\}$, which can be either 4 (all indices are distinct), 3 (one index is common to the two pairs), or 2 (both pairs of indices are the same).

- In the first case, we have independence of the two indicator random variables, and
$$\mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\mathbb{1}_{\{X_{s'}=X_{t'}\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\right]\mathbb{E}\left[\mathbb{1}_{\{X_{s'}=X_{t'}\}}\right] = \|\mathbf{p}\|_2^4.$$
- In the third case, the two indicator random variables are the same, and since $\mathbb{1}_{\{\}}^2 = \mathbb{1}_{\{\}}$ we get
$$\mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\mathbb{1}_{\{X_{s'}=X_{t'}\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\right] = \|\mathbf{p}\|_2^2.$$
- The second case is the messiest one; still, one can verify that in this case $\mathbb{1}_{\{X_s=X_t\}}\mathbb{1}_{\{X_{s'}=X_{t'}\}}$ is 1 if, and only if, the three distinct samples corresponding to the 3 distinct indices among s,t,s',t' take the same value, from which
$$\mathbb{E}\left[\mathbb{1}_{\{X_s=X_t\}}\mathbb{1}_{\{X_{s'}=X_{t'}\}}\right] = \|\mathbf{p}\|_3^3.$$

It remains to count how many summands of each type we have. Clearly, we have exactly $\binom{n}{2}$ summands of the third type; it is also not too hard to see that we have $\binom{n}{2}\binom{n-2}{2} = 6\binom{n}{4}$ summands of the first, and $6\binom{n}{3}$ of the second. (As a sanity check, $6\binom{n}{4} + 6\binom{n}{3} + \binom{n}{2} = \binom{n}{2}^2$, so all our summands are accounted for.)

[49] Namely, the summands $\mathbb{1}_{\{X_s=X_t\}}$ in the definition of Z are *positively correlated*:

$$\operatorname{Cov}(\mathbb{1}_{\{X_s=X_t\}}, \mathbb{1}_{\{X_{s'}=X_{t'}\}}) \ge 0$$

and are only independent if s,s',t,t' are all distinct.

Getting back to our variance computation, this yields

$$\begin{aligned}\mathrm{Var}[Z] &= \frac{1}{\binom{n}{2}^2}\left(6\binom{n}{4}\|\mathbf{p}\|_2^4 + 6\binom{n}{3}\|\mathbf{p}\|_3^3 + \binom{n}{2}\|\mathbf{p}\|_2^2\right) - \|\mathbf{p}\|_2^4 \\ &= \frac{1}{\binom{n}{2}^2}\left(\left(6\binom{n}{4} - \binom{n}{2}^2\right)\|\mathbf{p}\|_2^4 + 6\binom{n}{3}\|\mathbf{p}\|_3^3 + \binom{n}{2}\|\mathbf{p}\|_2^2\right) \qquad (79) \\ &\le \frac{4}{n}\|\mathbf{p}\|_3^3 + \frac{4}{n^2}\|\mathbf{p}\|_2^2 \\ &\le \frac{4}{n}\mathbb{E}[Z]^{3/2} + \frac{4}{n^2}\mathbb{E}[Z]\end{aligned}$$

first using that $6\binom{n}{4} < \binom{n}{2}^2$ to discard a negative term, then that $n \ge 2$ to get a simpler-looking upper bound on binomial coefficients, and finally writing $\|\mathbf{p}\|_3 \le \|\mathbf{p}\|_2$ by monotonicity of ℓ_p norms.

- In the case when $\mathbf{p} = \mathbf{u}_k$ (often called the *completeness* case), we need to control the probability that Z crosses our threshold $\tau := \frac{1+2\varepsilon^2}{k}$, that is

  $$\Pr[Z \ge \tau] = \Pr\left[Z \ge (1+2\varepsilon^2)\mathbb{E}[Z]\right] \le \Pr\left[Z \ge (1+\varepsilon^2)\mathbb{E}[Z]\right]$$

- in the "far" case (often called the *soundness* case), we want to control

  $$\Pr[Z < \tau] \le \Pr\left[Z < \frac{(1-\varepsilon^2)(1+4\varepsilon^2)}{k}\right] \le \Pr\left[Z < (1-\varepsilon^2)\mathbb{E}[Z]\right]$$

  using first that $(1-\varepsilon^2)(1+4\varepsilon^2) \ge 1+2\varepsilon^2$ (for $\varepsilon \le 1/2$), and then the fact that in the "far" case $\mathbb{E}[Z] > \frac{1+4\varepsilon^2}{k}$.

To control our probability of error in both cases, it is thus sufficient to upper bound $\Pr\left[|Z - \mathbb{E}[Z]| \ge \varepsilon^2\mathbb{E}[Z]\right]$; by Chebyshev's inequality (Theorem 11), this is at most

$$\begin{aligned}\Pr\left[|Z - \mathbb{E}[Z]| \ge \varepsilon^2\mathbb{E}[Z]\right] &\le \frac{\mathrm{Var}[Z]}{\varepsilon^4\mathbb{E}[Z]^2} \\ &\le \frac{4}{\varepsilon^4 n\mathbb{E}[Z]^{1/2}} + \frac{4}{\varepsilon^4 n^2\mathbb{E}[Z]} \\ &\le \frac{4\sqrt{k}}{\varepsilon^4 n} + \frac{4k}{\varepsilon^4 n^2}\end{aligned}$$

which is at most $1/3$, as desired, for $n \ge \frac{13\sqrt{k}}{\varepsilon^4}$. (For the third inequality, we relied on the fact that $\mathbb{E}[Z] = \|\mathbf{p}\|_2^2 \ge 1/k$ (cf. Remark 55.1).) □

The algorithm can be shown to work even for $n = O(\sqrt{k}/\varepsilon^2)$, but this require a much more careful variance analysis. As mentioned above, this $2/3$ can be boosted to $1-\delta$, for sample complexity $O\left(\frac{\sqrt{k}\log(1/\delta)}{\varepsilon^2}\right)$.

Yet we can do better! To conclude this lecture: we can do even better! The actual, optimal sample complexity of uniformity testing *has* been pinpointed,[50] and it is (perhaps surprisingly) a bit strange.

[50] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *ICALP*, volume 107 of *LIPIcs*, pages 41:1–41:14. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2018

Theorem 56. *Testing uniformity of an unknown distribution* $\mathbf{p} \in \Delta(k)$ *to total variation distance* ε *(with success probability* $1 - \delta$*) can be done with*

$$n = O\left(\frac{\sqrt{k\log(1/\delta)} + \log(1/\delta)}{\varepsilon^2}\right)$$

i.i.d. samples. (Moreover, this is optimal.)

The proof is outside the scope of this lecture, but note the rather strange dependence on δ! This is quite useful for very, very small δ.

A concluding remark. We may be tempted to consider a more robust version of testing: return yes when $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \leq \varepsilon_1$, and no when $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon_2$, for two arbitrary input parameters $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$. Unfortunately, this turns out to be a *much* harder problem, which (even when $\varepsilon_1, \varepsilon_2 = \Theta(1)$), requires $n = \Theta\left(\frac{k}{\log k}\right)$ samples![51]

[51] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new clts. In *Symposium on Theory of Computing Conference, STOC'11*, pages 685–694, 2011