Uniformity testing of discrete distributions

Clément L. Canonne

July, 2020

Abstract

The goal of this short note is to provide a short overview of the known algorithms to perform *uniformity testing* of discrete distributions over a known domain of size k.

The main focus of this document is the question of *uniformity testing* of probability distributions over a known discrete domain of size k[1]; that is, the question of deciding, based on observing a sequence of i.i.d. observations from some unknown probability distribution over domain $[k]$, whether this distribution is *the* uniform distribution $\mathbf{u}_k$ over the domain – or, in the contrary, is statistically quite far from this model. Formally, it is defined as follows, where

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) = \sup_{S\subseteq[k]}(\mathbf{p}(S)-\mathbf{q}(S)) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1 \in [0,1]$$

denotes the total variation distance between distributions:

Definition 1 (Uniformity Testing)**.** A *uniformity testing algorithm with sample complexity* n takes as input a parameter $\varepsilon \in (0,1]$ and n i.i.d. samples from an unknown distribution $\mathbf{p}$ over $[k]$, and outputs either accept or reject. The algorithm must satisfy the following, where the probability is over the randomness of the samples:

- If $\mathbf{p} = \mathbf{u}_k$, then the algorithm outputs accept with probability at least $2/3$;
- If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}_k) > \varepsilon$, then the algorithm outputs reject with probability at least $2/3$.

The sample complexity of uniformity testing is then the minimum sample complexity over all uniformity testing algorithms.

A couple remarks are in order: first, the above can be rephrased as a composite hypothesis testing (in a minimax setting), where $\mathcal{H}_0 = \{\mathbf{u}_k\}$ and $\mathcal{H}_1 = \{\ \mathbf{p} :\ \mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}_k) > \varepsilon\ \}$. Second, for simplicity, we focused in the above on a constant error probability (equal for both Type I and Type II), set to $1/3$. By standard arguments, one can in all settings considered here decrease this to an arbitrarily small $\delta \in (0,1]$ at the price of a mere multiplicative $\log(1/\delta)$ factor in the sample complexity,[2] by repeating the test independently and taking the majority outcome.

It is known [Pan08] that the sample complexity of uniformity testing with distance parameter $\varepsilon \in (0,1]$ is $\Theta(\sqrt{k}/\varepsilon^2)$. That's nice. Now, *how do we perform uniformity testing, though?* There are several things to consider in a testing algorithm. For instance:

Data efficiency: does the algorithm achieve the optimal sample complexity $\Theta(\sqrt{k}/\varepsilon^2)$?

Time efficiency: how fast is the algorithm to run (as a function of k, ε, and the number of samples n)?

Memory efficiency: how much memory does the algorithm require (as a function of k, ε, and n)?

[1] Without loss of generality, the set $[k] = \{1,2,\dots,k\}$

[2] Which is not optimal, as a $\sqrt{\log(1/\delta)}$ is achievable instead [DGPP18]; but is good enough.

Simplicity: is the algorithm simple to describe and implement?

"Simplicity": is the algorithm simple to *analyze*?

Robustness: how *tolerant* is the algorithm to breaches of the promise? I.e., does it accept distributions which are not *exactly* uniform as well, or is it very brittle?

Elegance: That's, like, your opinion, man.

Generalizable: Does the algorithm have other features that might be desirable in other settings?

Let's make a table, just with a couple of those criteria.

	Sample complexity	**Notes**	**References**
Collision-based	$\frac{k^{1/2}}{\varepsilon^2}$	Tricky	[GR00, DGPP19]
Unique elements	$\frac{k^{1/2}}{\varepsilon^2}$	$\varepsilon \gg 1/k^{1/4}$	[Pan08]
Modified χ^2	$\frac{k^{1/2}}{\varepsilon^2}$	Nope	[VV17, ADK15, DKN15]
Empirical distance to uniform	$\frac{k^{1/2}}{\varepsilon^2}$	Biased	[DGPP18]
Random binary hashing	$\frac{k}{\varepsilon^2}$	Fun	[ACT19]
Bipartite collisions	$\frac{k^{1/2}}{\varepsilon^2}$	$\varepsilon \gg 1/k^{1/10}$	[DGKR19]
Empirical subset weighting	$\frac{k^{1/2}}{\varepsilon^2}$	$\varepsilon \gg 1/k^{1/4}$	

Table 1: The current landscape of uniformity testing, based on the algorithms I know of. For ease of reading, we omit the $O(\cdot)$, $\Theta(\cdot)$, and $\Omega(\cdot)$'s from the table: all results should be read as asymptotic with regard to the parameters, up to absolute constants.

A key insight, that underlies a lot of the algorithms above, is that here ℓ_2 *distance is a good proxy for total variation distance*:

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

the inequality being Cauchy–Schwarz. So if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 > 4\varepsilon^2/k$ (and, well, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = 0$ then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 = 0$ too, of course). Moreover, we have the very convenient fact, specific to the distance to uniform: for any distribution $\mathbf{p}$ over $[k]$,

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

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

Collision-based. In view of the above, a very natural thing is to estimate $\|\mathbf{p}\|_2^2$, in order to distinguish between $\|\mathbf{p}\|_2^2 = 1/k$ (uniform) and $\|\mathbf{p}\|_2^2 > (1 + 4\varepsilon^2)/k$ (ε-far from uniform). How to do that? Upon observing that the probability that two independent samples x, y from $\mathbf{p}$ take the same value (a "collision") is exactly

$$\Pr_{x,y\sim\mathbf{p}}[\, x = y\,] = \sum_{i=1}^{k} \mathbf{p}(i)^2 = \|\mathbf{p}\|_2^2 \tag{3}$$

an obvious idea is to take n samples $x_1, \dots, x_n$, count the number of pairs that show a collision, and use that as an unbiased estimator Z_1 for $\|\mathbf{p}\|_2^2$:

$$Z_1 = \frac{1}{\binom{n}{2}} \sum_{s\neq t} \mathbb{1}_{\{x_s = x_t\}}\,. \tag{4}$$

By the above, $\mathbb{E}[Z_1] = \|\mathbf{p}\|_2^2$. If we threshold Z_1 at say $(1 + 2\varepsilon^2)/k$, we get a test. How big must n be for this to work? We can use Chebyshev for that, we requires to bound $\operatorname{Var}[Z]$. That's where things get tricky: to get the optimal bound $O(\sqrt{k}/\varepsilon^2)$ instead of an (easier to obtain) $O(\sqrt{k}/\varepsilon^4)$, the analysis of the variance has to be *pretty* intricate. Doable, but unwieldy.

Unique elements. Another idea? Count the number of elements that appear exactlly *once* among the n samples taken. Why is that a good idea? The uniform distribution will have the fewer collisions, so, equivalently, will have the maximum number of unique elements. In this case, the estimator Z_2 (the number of unique elements) has expectation

$$\mathbb{E}[Z_2] = n \sum_{i=1}^{k} \mathbf{p}(i)(1 - \mathbf{p}(i))^{n-1} \tag{5}$$

which is… a thing? Note that under the uniform distribution $\mathbf{u}_k$, this is exactly $n(1 - 1/k)^{n-1} \approx n - \frac{n^2}{k}$, and under arbitrary $\mathbf{p}$ this is (making a bunch of approximations not always valid) $\approx n \sum_{i=1}^{k} \mathbf{p}(i)(1 - n\mathbf{p}(i)) = n - n^2\|\mathbf{p}\|_2^2$. So the gap in expectation between the two cases "should" be around $4\varepsilon^2 n^2/k$, and, if the variance goes well and the stars align (and they do), we will be able to use Chebyshev and argue that we can distinguish the two for $n = \Theta(\sqrt{k}/\varepsilon^2)$.

Now, the annoying issue is that we count the number of *distinct* elements, and it's quite unlikely there can ever be more than k of them if the domain size is k. That explains, intuitively, the condition for the test to work: we need n (the number of samples taken) to be smaller than k (the maximum number of distinct elements one can ever hope to see), which gives, since we'll get $n = \Theta(\sqrt{k}/\varepsilon^2)$, the condition $\varepsilon \gg 1/\varepsilon^{1/4}$. (A slight bummer.)

Modified χ^2. If you are a statistician, or just took Stats 101, or just got lost on Wikipedia at some point and randomly ended up on the wrong page, you may know of Pearson's χ^2 test for goodness-of-fit: for every element i of the domain, count how many times it appeared in the samples, N_i. Compute $\sum_i \frac{(N_i - n/k)^2}{n/k}$. Relax.

The bad news is that it does not actually lead to the optimal sample complexity: the variance of this thing can be too big, due to the elements we only expect to see zero or once (so, most of them). The *good* news is that a simple correction of that test, of the form

$$Z_3 = \sum_{i=1}^{k} \frac{(N_i - n/k)^2 - N_i}{n/k} \tag{6}$$

does have a much smaller variance, and a threshold test of the form "$Z_3 > \tau$?" leads to the right sample complexity. To understand why, it's helpful to think of taking $\operatorname{Poisson}(n)$ samples instead of exactly n, as it simplifies the variance analysis (and changes roughly nothing else here). Then the N_i's become independent, with $N_i \sim \mathrm{Poisson}(n\mathbf{p}(i))$ (that not magic, it's Poissonization). The expectation of Z_3 is then just

$$\mathbb{E}[Z_3] = nk\|\mathbf{p} - \mathbf{u}_k\|_2^2$$

which is perfect. Analyzing this test just boils down, again, to bounding the variance of Z_3 and invoking Chebyshev's inequality… It's a good exercise, and under the Poissonization assumption not that hard. (Try *without* removing N_i in the numerator, though, and see what you get…)

Empirical distance to uniform. Let's take a break from ℓ_2 and consider another, very natural thing to consider: the *plugin estimator*. Since we have n samples from $\mathbf{p}$, we can compute the empirical estimator of the distribution, $\hat{\mathbf{p}}$. Now, we want to test whether $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = 0$ v. $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$? Why not consider

$$Z_4 = \mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathbf{u}_k) \tag{7}$$

the empirical distance to uniform? A reason might be: *this sounds like a terrible idea.* Unless $n = \Omega(k)$ (which is much more than what we want), the empirical distribution $\hat{\mathbf{p}}$ will be at distance $1 - o(1)$ from uniform, *even* if $\mathbf{p}$ is actually uniform.

That's the thing, though: hell is in the $o(1)$ details. Sure, $\mathbb{E}[Z_4]$ will be *almost* 1 whether $\mathbf{p}$ is uniform or far from it unless $n = \Omega(k)$. But this "almost" will be different in the two cases! Carefully analyzing this tiny gap in expectation, and showing that Z_4 concentrates well enough around its expectation to preserve this tiny gap, amazingly leads to a tester with optimal sample complexity $n = \Theta(\sqrt{k}/\varepsilon^2)$.

Random binary hashing. Now for a tester that is *not* sample-optimal (but has other advantages, and is relatively cute). If there is one thing we know how to do optimally, it's estimating the bias of a coin. We don't have a coin (Bernoulli) here, we have a glorious $(k-1)$-dimensional object. Hell, let's just randomly make it a coin, shall we? Pick your favourite (4-wise independent) hash function $h\colon [k] \to \{0, 1\}$, thus randomly partitioning the domain $[k]$ in two sets S_0, S_1. Hash all the n samples you got: *now* we have a random coin!

Let's estimate its bias then: we know exactly what this should be under the uniform distribution: $\mathbf{u}_k(S_0)$. If only we could argue that $\mathbf{p}(S_0)$ noticeably differs from $\mathbf{u}_k(S_0)$ (with high probability over the random choice of the hash function) whenever $\mathbf{p}$ is ε-far from uniform, we'd be good. Turns out… it is the case:

$$\Pr_{S \subseteq [k]}\left[\, |\mathbf{p}(S) - \mathbf{u}_k(S)| = \Omega(\varepsilon/\sqrt{k}) \,\right] = \Omega(1) \tag{8}$$

So we can just do exactly this: we need to estimate the bias $\mathbf{p}(S_0)$ up to an additive $\alpha \asymp \varepsilon/\sqrt{k}$. This can be done with $n = \Theta(1/\alpha^2) = \Theta(k/\varepsilon^2)$ samples, as desired.

Bipartite collisions. In the collision-based tester above, we took a multiset S of n samples from $\mathbf{p}$, and looked at the number of "collisions" in S to define our statistic Z_1. That is fine, but requires to keep in memory all the samples observed so far. One related idea would be to instead take *two* multisets S_1, S_2 of n_1 and n_2 samples, and only count "bipartite collisions," i.e., collisions between a sample of S_1 and one of S_2:

$$Z_5 = \frac{1}{n_1 n_2} \sum_{(x,y) \in S_1 \times S_2} \mathbb{1}_{\{x=y\}} \tag{9}$$

One can check that $\mathbb{E}[Z_5] = \|\mathbf{p}\|_2^2$. Back to ℓ_2 as proxy! Compared to the "vanilla" collision-based test, this is more flexible (S_1, S_2 need not be of the same size), and thus lends itself to some settings where a tradeoff between n_1 and n_2 is desirable (roughly speaking, one needs $n_1 n_2 \gtrsim k/\varepsilon^4$, and the sample complexity is $n = n_1 + n_2$). For the case $n_1 = n_2$, this retrieves the optimal $n \asymp \sqrt{k}/\varepsilon^2$, with some extra technical condition stemming from the analysis, unfortunately: one needs $\varepsilon = \Omega(1/k^{1/10})$.

Empirical subset weighting. That one, I really like. It's adaptive, it's weird, and (I think) it's new. Fix a parameter $1 \leq s \leq n$. Take n samples from $\mathbf{p}$, and consider the set S (not multiset) induced by the first s samples you get. One can check that

$$\mathbb{E}[\mathbf{p}(S)] = \sum_{i=1}^{k} \mathbf{p}(i)(1-(1-\mathbf{p}(i))^s) \tag{10}$$

which should be roughly (making a bunch of approximations) $\mathbb{E}[\mathbf{p}(S)] \approx s\|\mathbf{p}\|_2^2$. Under the uniform distribution, this is exactly $(1-(1-1/k)^s) \approx s/k$, where the approximation is valid for $s \ll k$.

Great: we have a new estimator for (roughly) the ℓ_2 norm! Now, assuming things went well, as the end of this first stage we have a set S such that $\mathbf{p}(S)$ is approximately either s/k or $s\|\mathbf{p}\|_2^2 \geq s(1+\Omega(\varepsilon^2))/k$ (we just argued that this is what things happen *in expectation*).[3] So, let's do a second stage! Take the next $n-s$ samples, and just count the number of them which fall in S: this is allows you to estimate $\mathbf{p}(S)$ up to an additive $s\varepsilon^2/k$, as long as

$$n - s \gtrsim \frac{k}{s\varepsilon^4}$$

(exercise: check that). Optimizing, we get that for $s = n/2$ this leads to $n \asymp \sqrt{k}/\varepsilon^2$: optimal sample complexity! Only drawback: we need $s \ll k$ for our approximations to be valid (after that, $\mathbb{E}[\mathbf{p}(S)]$ cannot be approximately $s\|\mathbf{p}\|_2^2$ anymore; same issue as with the "unique elements" algorithm), so we get the condition $\varepsilon \gg 1/k^{1/4}$. Sight bummer.

References

[ACT19] Jayadev Acharya, Clement Canonne, and Himanshu Tyagi. Communication-constrained inference and the role of shared randomness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 30–39, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *CoRR*, abs/1507.05952, 2015.

[DGKR19] Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, and Sankeerth Rao. Communication and memory efficient testing of discrete distributions. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 1070–1106. PMLR, 2019.

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 41, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

[DGPP19] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chic. J. Theoret. Comput. Sci.*, pages Art. 1, 21, 2019.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *SODA*, pages 1841–1854. SIAM, 2015.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[3] Some more details are required to argue $\mathbf{p}(S)$ does concentrate enough around its expectation.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014*, pages 51–60. IEEE Computer Soc., Los Alamitos, CA, 2014.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. Journal version of [VV14].