The Price of Tolerance in Distribution Testing

Clément L. Canonne CLEMENT.CANONNE@SYDNEY.EDU.AU
University of Sydney

Ayush Jain AYJAIN@ENG.UCSD.EDU
UC San Diego

Gautam Kamath G@CSAIL.MIT.EDU
Cheriton School of Computer Science, University of Waterloo

Jerry Li JERRL@MICROSOFT.COM
Microsoft Research

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We revisit the problem of tolerant distribution testing. That is, given samples from an unknown distribution p over $\{1,\dots,n\}$, is it ε_1-close to or ε_2-far from a reference distribution q (in total variation distance)? Despite significant interest over the past decade, this problem is well understood only in the extreme cases. In the noiseless setting (i.e., $\varepsilon_1 = 0$) the sample complexity is $\Theta(\sqrt{n})$, strongly sublinear in the domain size. At the other end of the spectrum, when $\varepsilon_1 = \varepsilon_2/2$, the sample complexity jumps to the barely sublinear $\Theta(n/\log n)$. However, very little is known about the intermediate regime. We fully characterize the price of tolerance in distribution testing as a function of n, ε_1, ε_2, up to a single $\log n$ factor. Specifically, we show the sample complexity to be

$$\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n}\cdot\max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right),$$

providing a smooth tradeoff between the two previously known cases. We also provide a similar characterization for the problem of tolerant equivalence testing, where both p and q are unknown. Surprisingly, in both cases, the main quantity dictating the sample complexity is the ratio $\varepsilon_1/\varepsilon_2^2$, and not the more intuitive $\varepsilon_1/\varepsilon_2$. Of particular technical interest is our lower bound framework, which involves novel approximation-theoretic tools required to handle the asymmetry between ε_1 and ε_2, a challenge absent from previous works.

Keywords: distribution testing, sample complexity, algorithms, lower bounds, hypothesis testing

1. Introduction

Upon observing independent samples from an unknown probability distribution, can we determine whether it possesses some property of interest? This natural question, known as distribution testing or statistical hypothesis testing, has enjoyed significant study from several communities, including theoretical computer science, statistics, information theory, and machine learning. The prototypical problem in this area is *identity testing* (sometimes called *goodness-of-fit* or *one-sample* testing): given samples from an unknown probability distribution p over $[n]$, test whether it is equal to some reference distribution q, or ε-far in ℓ_1-distance. It is now well understood that $\Theta\left(\sqrt{n}/\varepsilon^2\right)$ samples are necessary and sufficient to solve this problem (Ingster, 1994; Goldreich and Ron, 2000; Batu et al., 2001; Paninski, 2008; Valiant and Valiant, 2014; Diakonikolas et al., 2015; Acharya et al., 2015; Diakonikolas et al., 2018). Quite surprisingly, this sample complexity is strongly sublinear in n, enabling sample-efficient testing even over large domains.

The drawback of this formulation is that it is very particular in terms of the relationship between p and q. More precisely, it prescribes only that one must distinguish between the cases where p and q are far versus when they are *exactly* equal – no guarantees are provided for any intermediate case, e.g., for when p and q are *close* but not identical. This restriction limits the relevance of solutions to this problem, as it is unrealistic to assume *precise* knowledge of a distribution due to a number of reasons, including model misspecification, imprecise measurements, or dataset contamination.
To address these concerns, the problem of *tolerant identity testing* was introduced (Parnas et al., 2006), which is the main focus of our work.

> *Tolerant Identity Testing*: Given an explicit description of a distribution q over $[n]$, sample access to a distribution p over $[n]$, and bounds $\varepsilon_2 > \varepsilon_1 \geq 0$, and $\delta > 0$, distinguish with probability at least $1-\delta$ between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$, whenever p satisfies one of these two inequalities.

We will also study the problem of *tolerant equivalence testing* (sometimes called tolerant *closeness* or *two-sample* testing):

> *Tolerant Equivalence Testing*: Given sample access to distributions p and q over $[n]$, and bounds $\varepsilon_2 > \varepsilon_1 \geq 0$, and $\delta > 0$, distinguish with probability at least $1-\delta$ between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$, whenever p, q satisfy one of these two inequalities.

Focusing our attention on tolerant identity testing and constant ε_2, it is natural to consider the strong tolerance requirement of $\varepsilon_1 = \varepsilon_2/2$, in which the two cases are separated only by a constant factor. One would ideally like to maintain the strongly sublinear sample complexity of $\mathcal{O}(\sqrt{n})$, as in the non-tolerant case where $\varepsilon_1 = 0$. Unfortunately, this is impossible: as shown by Valiant and Valiant (2010a,b, 2011a), $\Theta\big(\frac{n}{\log n}\big)$ samples are necessary and sufficient, see also (Jiao et al., 2018, 2017; Han et al., 2016). On the other end of the spectrum, it is known that mild tolerance of $\varepsilon_1 = \frac{\varepsilon_2}{2\sqrt{n}}$ is achievable with the same strongly-sublinear sample complexity of $\mathcal{O}(\sqrt{n})$, by converting ℓ_2-tolerance to ℓ_1-tolerance (Goldreich and Ron, 2000; Batu et al., 2001, 2013; Diakonikolas et al., 2015; Diakonikolas and Kane, 2016; Daskalakis et al., 2018). However, existing results only capture these two extremes, and we have very little understanding of the intermediate landscape of tolerant testing. Does there exist a smooth hierarchy of increasingly difficult testing problems, or is there a sharp transition in the sample complexity from strongly to barely sublinear?

1.1. Results and Techniques

We provide a complete characterization of the sample complexity of tolerant identity and equivalence testing (up to a single logarithmic factor in the domain size n). Our main results are as follows:

Theorem 1 (Identity testing (Informal; see Theorem 5 and Corollary 8)) *The sample complexity of tolerant identity testing over $[n]$ with parameters $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ is*

$$\Omega\left(\frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n}{\log n} \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right) \text{ and } \mathcal{O}\left(\frac{\sqrt{n}}{\varepsilon_2^2} + n \cdot \max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right).$$

Theorem 2 (Equivalence testing (Informal; see Theorem 6 and Corollary 9)) *The sample complexity of tolerant equivalence testing over $[n]$ with parameters $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ is*

$$\Omega\left(\max\left\{\frac{\sqrt{n}}{\varepsilon_2^2}, \frac{n^{2/3}}{\varepsilon_2^{4/3}}\right\} + \frac{n}{\log n}\cdot\max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right) \textit{ and } \mathcal{O}\left(\max\left\{\frac{\sqrt{n}}{\varepsilon_2^2}, \frac{n^{2/3}}{\varepsilon_2^{4/3}}\right\} + n\cdot\max\left\{\frac{\varepsilon_1}{\varepsilon_2^2}, \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right\}\right).$$

In both cases, we give computationally-efficient algorithms which achieve the upper bounds. Moreover, one interesting feature of our algorithms is that they require no knowledge of ε_1, which only arises in the sample complexity: that is, our algorithm automatically achieves the best possible ε_1, for a given target ε_2 and number of samples.

It is worth noting that prior to our work, only two extreme points of the full tradeoff we show were known:

- the "non-tolerant" (noiseless) case where $\varepsilon_1 = 0$, for which the $\Theta(\sqrt{n}/\varepsilon_2^2)$ sample complexity (or, for the equivalence testing version, $\Theta(\max\{\sqrt{n}/\varepsilon_2^2, n^{2/3}/\varepsilon_2^{4/3}\})$) (Paninski, 2008; Valiant, 2011; Chan et al., 2014; Valiant and Valiant, 2014). In the case of identity testing, it is further known that some of the optimal testers (namely, those based on testing in the ℓ_2 distance as a proxy) achieve a weak tolerance of $\varepsilon_1 = \varepsilon_2/\sqrt{n}$ "for free", due to the relation between ℓ_1 and ℓ_2 norm along with the Cauchy–Schwarz inequality.

- the maximally noisy case where $\varepsilon_1 = \Theta(\varepsilon_2)$, for which results of Valiant and Valiant (2010a,b, 2011a) as well as follow-up works (Jiao et al., 2018, 2017; Han et al., 2016) show that the sample complexity must grow as $\Theta(n/\log n)$. Interestingly, the dependence on $\varepsilon_1, \varepsilon_2$ was not fully understood, even in this case, as most lower bounds dealt with *estimation* of the distance between p, q to an additive ε, which is a related yet different problem (essentially, showing that $\Omega(n/(\varepsilon_2 - \varepsilon_1)^2 \log n)$ samples are required, when $\varepsilon_1 = \Theta(1)$ and $\varepsilon_2 - \varepsilon_1$ can be arbitrarily small). The lower bound from (Valiant and Valiant, 2010a) does imply, by "scaling," an $\Omega(n/(\varepsilon_2 \log n))$ lower bound for arbitrary ε_2 and $\varepsilon_1 = \Theta(\varepsilon_2)$, but it is still far from the upper bound of $\mathcal{O}(n/(\varepsilon_2^2 \log n))$ in this regime that both (Valiant and Valiant, 2010b) and (Jiao et al., 2018) prove in this setting. Our result shows that this upper bound is tight in this parameter regime, as our lower bound is then $\Omega(n/(\varepsilon_2^2 \log n))$.

We emphasize that our results go beyond those two extreme points, and essentially settles the landscape of tolerant testing. As just one example, the question of testing $1/n^{1/10}$-close vs. $1/n^{1/5}$-far was left completely open by previous work; our results imply that the sample complexity is $\tilde{\Theta}(n)$. We depict in Figure 1 the different regimes of sample complexity this leads to, for both identity and closeness testing.

Surprisingly, our results for both tolerant identity and equivalence testing show that the relevant quantity governing the "price of tolerance" is not the ratio $\varepsilon_1/\varepsilon_2$, as one might naïvely think; but instead is the (inhomogeneous!) ratio $\rho \coloneqq \varepsilon_1/\varepsilon_2^2$, which might seem counterintuitive – especially in view of the two different regimes the $\max(\rho, \rho^2)$ scaling implies.

Another interesting and unexpected byproduct of our result is to show that even the known "weak tolerance" of the standard ℓ_2-based testers, which allow to test identity with tolerance $\varepsilon_1 = \varepsilon_2/\sqrt{n}$ with the same $\mathcal{O}(\sqrt{n}/\varepsilon_2^2)$ sample complexity as the non-tolerant case, is *not* the best one can do with this sample complexity. Indeed, our results imply that one can actually achieve tolerance up to $\varepsilon_1 = \min(1/\sqrt{n}, \varepsilon_2/\sqrt[4]{n})$ "for free," a significant improvement over $\varepsilon_2/\sqrt{n}$. One can rephrase

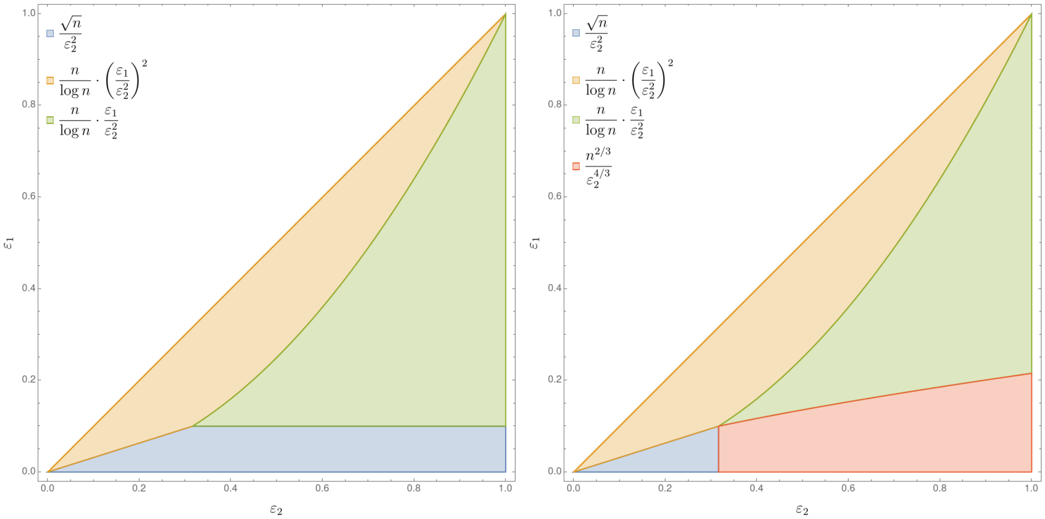

Figure 1: The different regimes of sample complexity corresponding to Theorem 1 (identity testing, left) and Theorem 2 (closeness testing, right), as a function of $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ (for fixed n), depicting in both cases which of the terms of the sample complexity bound dominates.

this as saying that the Cauchy–Schwarz inequality, from which this "natural" weak tolerance provided by ℓ_2-based testers stems from, is (oddly) not the right way to look at the problem.

Finally, our techniques allow us to derive an analogue of Theorem 1 for the "instance-optimal" setting (Valiant and Valiant, 2017) (see also (Blais et al., 2017; Diakonikolas and Kane, 2016)), where the sample complexity is expressed as a function of the known reference distribution q instead of the domain size n (which corresponds to a worst-case over all possible reference distributions). Specifically, we show the following:[1]

Theorem 3 (Instance-optimal identity testing (Informal; see Theorem 39 and Theorem 38)) *For any fixed q over $\mathbb{N}$, the sample complexity of tolerant identity testing with reference distribution q with parameters $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ is*

$$\tilde{\Theta}\left(\frac{\|q_{-\Theta(\varepsilon_2)}\|_{2/3}}{\varepsilon_2^2} + \|q_{-\Theta(\varepsilon_2)}\|_{1/2} \cdot \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \|q_{-\Theta(\varepsilon_2)}\|_0 \cdot \frac{\varepsilon_1}{\varepsilon_2^2} \right),$$

where $q_{-\alpha}$ denotes the (sub)distribution obtained by removing as many of the smallest elements of q as possible, without removing more than a total of α probability mass overall.

We defer the details and proof of this result to section D; and discuss some of its aspects here. First, note that by choosing q to be the uniform distribution, we see that $\|q_{-\Theta(\varepsilon_2)}\|_{2/3} \approx \sqrt{n}$, $\|q_{-\Theta(\varepsilon_2)}\|_{1/2} \approx n$, and $\|q_{-\Theta(\varepsilon_2)}\|_0 \approx n$, so that Theorem 3 retrieves Theorem 1 up to logarithmic factors. In particular, this gives a refined perspective on Theorem 1, showing that the $\tilde{\Theta}(n)$ term actually arises due to two separate costs, which happen to coincide for the uniform distribution.

[1] Here and in section D, we slightly abuse the $\tilde{\Theta}$ notation to also hide logarithmic factors in n, not just in the argument.

This brings us to our second point: the term $\|q_{-\Theta(\varepsilon_2)}\|_{2/3}$ corresponds to the *non-tolerant* instance-optimal identity testing bound established in (Valiant and Valiant, 2017), i.e., a *testing* term; while the quantity $\|q_{-\Theta(\varepsilon_2)}\|_{1/2}$ can be interpreted as capturing the difficulty of *learning*, as the $1/2$-quasinorm is known to capture the sample complexity of learning a probability distribution (see, e.g., (Kamath et al., 2015; Canonne, 2020b)). Thus, the bound of Theorem 3 can be read as saying the sample complexity of tolerant identity testing is (nearly) characterized by three aspects of the reference distribution: how hard it is to *test*, how hard it is to *learn*, and how large its *effective support size* is.

Relation to the Statistics literature. Despite their pervasive use, the Statistics community is outspoken about the pitfalls associated with point nulls (i.e., $\varepsilon_1 = 0$) for statistical hypothesis testing (Berger and Sellke, 1987; Rao and Lovric, 2016; Abadie, 2020). Instead, the community advocates for composite nulls, where the null hypothesis is a set of distributions rather than a single one. This more-general problem is often reduced to our tolerant testing problem (sometimes called the *imprecise null* by the Statistics community) by assuming the null holds and performing estimation to obtain a candidate distribution q. Thus, we believe our results may be a useful tool for solving more challenging composite-versus-composite hypothesis testing problems. While some classic work provides minimax rates for certain related tolerant testing problems (Ingster, 2000), results in this direction have been relatively hard to come by. A recent survey paper by Balakrishnan and Wasserman (2018), specifically highlight the problem of designing non-conservative thresholds for imprecise null hypothesis tests, which we believe to be an interesting direction for future work.

Overview of our techniques. Given the extensive literature on distribution testing, the community has developed a rich set of tools for problems in this space. However, the techniques used for the two extreme cases appear to be qualitatively quite different. In the non-tolerant case, algorithms usually take the form of simple ℓ_2- or χ^2-test statistics, and lower bounds are established via either Ingster's method (Ingster, 1994) or mutual information arguments. On the other hand, analysis for the maximally noisy case depends on results from the literature on best-polynomial approximation. Given the contrasting approaches for these two cases, it is natural to wonder which set of techniques will be effective for the problems which lie between the two. Interestingly, our results borrow from both: our algorithms are more similar to those from the non-tolerant setting, while our lower bound techniques resemble those in the maximally noisy case.

Our main algorithm thresholds a rescaled ℓ_2-statistic (in certain cases called a χ^2-statistic), similar to testing algorithms in the past (see, e.g., (Chan et al., 2014; Valiant and Valiant, 2014; Acharya et al., 2015; Diakonikolas et al., 2015)). Specifically, our statistic takes the form $Z = \sum_i \left((X_i - Y_i)^2 - X_i - Y_i\right)/\hat{f}_i$, where X_i and Y_i are the number of occurrences of symbol i drawn from distribution p and q, respectively, and $\hat{f}_i$ are symbol-dependent rescaling factors. While prior works either computed these factors for identity testing based on the reference distribution q, or used the same set of samples for both the numerator as well as the rescaling factor in the denominator, we use sample-splitting to separately obtain empirical estimates $\hat{f}_i$ for the relevant quantities. We show multiplicative concentration for these factors to ensure they are close to the values for which we are using them as a proxy. These rescaling factors are empirical estimates of two terms. A typical choice, now common in the literature, is based on $p_i + q_i$, which limits fluctuations in the estimator caused by individual terms. Our approach crucially introduces an additional novel rescaling term based on $|p_i - q_i|$, which prevents the statistic from placing too much emphasis on the the ℓ_2-norm of the distribution. The contributions of both terms are crucial for making the analysis work out.

We note that our test statistic only involves the first two moments of the distribution. This is in contrast to previous upper bounds for tolerant testing in the maximally noisy case, which instead inspected $\log n$ moments. Thus, we show that considering only two moments suffices for near-optimal tolerant testing. Interestingly, our algorithm achieves the optimal sample complexity (up to constants) for the non-tolerant case, but loses a $\log n$ factor in the maximally noisy case. Removing this final logarithmic term may require a statistic which exploits higher-order moments, and is an interesting question for future work.

Our lower bounds are obtained via the generalized two-point method. At a high level, we follow the moment matching approach pioneered by Wu and Yang (2016). We construct two priors over distributions, where distributions drawn from the two priors are ε_1-close to and ε_2-far from uniform, respectively. To prove lower bounds, we must choose these priors such that the process of drawing a distribution and then m samples from it has low total variation distance between the two priors. By considering priors over product distributions, we can further reduce our task to simply constructing a pair of univariate random variables with properties described in Theorem 7. By appealing to results from polynomial approximation (Lemma 11), it suffices to construct this pair such that their low-order moments match.

Prior works construct this pair of random variables by expressing this moment matching problem as an infinite dimensional convex program and analyzing its dual. Our approach follows the same recipe, however, the analysis of the dual convex program is much more involved in our case. Prior lower bounds only considered the special case where ε_1 and ε_2 differ by a fixed, constant factor. In this regime, tolerant testing becomes essentially equivalent to learning the ℓ_1 distance between p to q to error ε_1. This is a setting which is much easier for this formalism to handle; indeed, the moment matching paradigm was initially designed for estimation problems. Importantly, this induces a key symmetry in the lower bound construction, and consequently, the dual has a very nice interpretation in terms of the best uniform approximation of a given function by a low-degree polynomial.

In our case we must handle general ε_1 and ε_2, and this symmetry is lost. As a result, we analyze a convex program which directly captures the testing problem. However, the dual has a much more complex interpretation. At a high level, the goal is now to approximately fit a low-degree polynomial within a “wedge” of minimal arc length. Interestingly, in our formulation of the dual, instead of having to prove that there is a good approximating polynomial, we must demonstrate that no low degree polynomial can achieve this task. This is the main technical difficulty in the lower bound, and we do so from first principles by leveraging classic tools from polynomial approximation theory to prove new approximation-theoretic results in our setting. Due to space constraints, all proofs are deferred to the appendix, and this extended abstract focuses on providing an outline of the results and arguments.

1.2. Related Work

Distribution testing was first considered in the theoretical computer science community by Goldreich and Ron (2000), who analyzed and applied an algorithm for uniformity testing towards the problem of testing whether a graph is an expander. Batu, Fischer, Fortnow, Kumar, Rubinfeld, and White (Batu et al., 2001) studied the general problem of identity testing. A number of results have discovered and rediscovered optimal bounds for identity testing (Paninski, 2008; Valiant and Valiant, 2014; Acharya et al., 2015; Diakonikolas et al., 2015; Goldreich, 2016; Diakonikolas and

Kane, 2016; Diakonikolas et al., 2018; Daskalakis et al., 2018; Diakonikolas et al., 2019), even with optimal dependence on the failure probability δ and on an instance-by-instance basis. The harder problem of equivalence testing was studied in (Batu et al., 2000), and optimal upper and lower bounds were given in (Valiant, 2011; Chan et al., 2014; Daskalakis et al., 2018; Diakonikolas et al., 2021). Some work has also studied the case where an unequal number of samples are received from the two distributions (Acharya et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016).

Tolerant testing has been previously considered, in a few different regimes. Strong tolerance, or equivalently, estimating distance between distributions, was studied first by Valiant and Valiant (2010a,b, 2011a,b), and in more recent works by Han, Jiao, Venkat, and Weissman (Jiao et al., 2018, 2017; Han et al., 2016). Tolerance in distances besides ℓ_1 (including chi-squared, KL, Hellinger, and ℓ_2) has also been considered (Goldreich and Ron, 2000; Batu et al., 2000; Chan et al., 2014; Acharya et al., 2015; Daskalakis et al., 2018). An interesting direction for future work is to understand the sample complexity of tolerant testing for these other distances in a fine-grained manner, as we do for ℓ_1 distance. Moreover, results with ℓ_2-tolerance imply testers with weak ℓ_1-tolerance, through the relation between ℓ_1 and ℓ_2 norms and the Cauchy–Schwarz inequality. Finally, very recent work sets out to understand whether, for general properties of distributions, the (near)-quadratic gap between tolerant and non-tolerant testing achievable for identity testing is the worst possible (Chakraborty et al., 2021). For additional background on distribution testing, see surveys and related work in (Rubinfeld, 2012; Balakrishnan and Wasserman, 2018; Kamath, 2018; Canonne, 2020a).

Techniques involving moment matching and best-polynomial approximation are useful for tolerant distribution testing, but also play a key role in estimation of distributional properties, including entropy, support size, support coverage, and distance to uniformity (Raskhodnikova et al., 2009; Wu and Yang, 2016; Jiao et al., 2018; Acharya et al., 2017; Orlitsky et al., 2016; Wu and Yang, 2018) See (Wu and Yang, 2020) for a survey on applications of polynomial methods in statistics.

Preliminaries. We identify a probability distribution p over a known discrete domain $[n] := \{1, 2, \ldots, n\}$ with its probability mass function (pmf), i.e., a nonnegative vector $p = (p_1, p_2, \ldots, p_n)$ such that $\sum_{i=1}^n p_i = 1$. Given two distributions p, q, their total variation distance (also known as statistical distance) is defined as

$$\mathrm{TV}(p,q) = \sup_{S\subseteq[n]} (p(S) - q(S)) = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i| = \frac{1}{2}\|p - q\|_1 .$$

Due to this equivalence with the ℓ_1 norm, we will interchangeably use the TV and ℓ_1 norms in our paper. We will also extensively use the ℓ_2 distance between probability distributions, which is just the ℓ_2 norm $\|p - q\|_2$ between their pmfs and, by Cauchy–Schwarz, satisfies $\frac{1}{\sqrt{n}}\|p - q\|_1 \leq \|p - q\|_2 \leq \|p - q\|_1$.

Let p, q be two distributions over the domain $[n]$. For given ε_1 and ε_2 such that $0 \leq \varepsilon_1 < \varepsilon_2$, we want to understand the sample complexity (i.e., minimum number of i.i.d. samples required) to distinguish between:

Yes: $\|p - q\|_1 \leq \varepsilon_1$,

No: $\|p - q\|_1 > \varepsilon_2$.

with probability at least $4/5$.[2]

We consider the problems of tolerant uniformity, identity, and equivalence testing. In *identity testing* the distribution q is explicitly known in advance, while p is unknown: the sample complexity is then the number of i.i.d. samples from p. *Uniformity testing* is a special case of identity testing, where $q = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$ is the uniform distribution, denoted Unif_n. In *equivalence testing*, both p and q are unknown and we get samples from each. The sample complexity is then the total number of samples obtained from both p and q. We will typically denote the number of i.i.d. samples used by an algorithm by m. Note that uniformity testing is a special case of identity testing (and hence lower bounds for the former imply lower bounds for the latter), and that equivalence testing is at least as hard as identity testing, in terms of sample complexity.

2. Algorithms for Tolerant Testing

In this section, we describe our testing algorithm (Algorithm 1), before analyzing its performance. As a preliminary simplification, instead of assuming the algorithm is provided with m independent samples we will rely on the so-called "Poissonization trick" and assume we obtain $\mathrm{Poi}(m)$ samples each from both p and q. The benefit of Poissonization is that the number of occurrences of each domain element will be an independent Poisson, eliminating correlations between symbols which arise with a fixed budget. This is without loss of generality, as by standard arguments about concentration of Poisson random variables this changes the sample complexity by at most a (small) constant factor. Moreover, losing again a factor 2 in the sample complexity, our algorithms will take as input two sets of $\mathrm{Poi}(m)$ samples for each of p and q.

Let $\tilde{X}_i$ and X_i be the count of occurrences of symbol $i \in [n]$ in the first and the second set of the samples from p, respectively. Similarly, let $\tilde{Y}_i$ and Y_i be the count of symbol i in the first and the second set of the samples from q, respectively. Let

$$f_i := \begin{cases} \max\{\sqrt{mn} \cdot |p_i - q_i|, n \cdot (p_i + q_i), 1\} & \text{if } m \geq n \\ \max\{m \cdot (p_i + q_i), 1\} & \text{if } m < n. \end{cases}$$

We will use the first set of counts $\tilde{X}_i$ and $\tilde{Y}_i$ to estimate f_i with $\widehat{f_i}$, defined as

$$\widehat{f_i} := \begin{cases} \max\left\{ \frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}}, \frac{\tilde{X}_i + \tilde{Y}_i}{m/n}, 1 \right\} & \text{if } m \geq n \\ \max\{\tilde{X}_i + \tilde{Y}_i, 1\} & \text{if } m < n. \end{cases}$$

Let $Z_i := (X_i - Y_i)^2 - X_i - Y_i$ and let

$$Z := \sum_{i=1}^{n} \frac{Z_i}{\widehat{f_i}}. \tag{1}$$

and $\tau := c \cdot \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2 \varepsilon_2^2}{n}\right)$, where $c > 0$ is an absolute constant determined in the course of the analysis. Our tester is then as follows: Note that the algorithm itself requires no knowledge of ε_1,

[2] The exact constant here is immaterial, and by standard amplification arguments one can achieve a probability of success of $1 - \delta$ at the cost of a multiplicative $\mathcal{O}(\log(1/\delta))$ factor in the sample complexity.

```
Algorithm 1: Tolerant testing algorithm.
Data: $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, $m$, $n$, two sets of $\mathrm{Poi}(m)$ samples from both $p$ and $q$
Set the threshold
    $\tau \leftarrow c \cdot \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$
Compute $Z$ from the sets of samples, as per (1)
if $Z \geq \tau$ then return $\|p - q\|_1 \geq \varepsilon_2$
return $\|p - q\|_1 \leq \varepsilon_1$
```

and thus as the number of samples m increases, the same test statistic (with appropriate substitution of m) becomes more and more tolerant.

To gain some intuition, we first remark that our tester is a modification of the ℓ_2 testers in (Chan et al., 2014; Diakonikolas and Kane, 2016), and akin to the chi-square tester from (Acharya et al., 2015). The main difference lies in the choice of normalizing factor f_i (of which $\widehat{f_i}$ is merely the natural estimator). The goal of this denominator is twofold: the relatively standard term $n \cdot (p_i + q_i)$ (which is comparable to the standard deviation of Z_i; for $m < n$, we use $m(p_i + q_i)$ to make up for larger imprecision in our estimates) ensures that no single term of the sum will make the estimator fluctuate too much. The term $\sqrt{mn} \cdot |p_i - q_i|$ (which is only needed the regime $m \geq n$, since for $m < n$ the best accuracy we can get for $|p_i - q_i|$ is $\approx 1/m$, but scaling by $m|p_i - q_i|$ would be unnecessary as the $m(p_i + q_i)$ term already dominates) is a crucial difference with previous work; its goal is to "tamper down" the numerator Z_i when the ℓ_2 contribution $(p_i - q_i)^2$ is too large, which is key for our ℓ_2-based tester to work. Indeed, in the "far" case where $\|p - q\|_1 \geq \varepsilon_2$, this is not a problem; however, in the "close" case where $\|p - q\|_1 \leq \varepsilon_1$, the relation between ℓ_2 and ℓ_1 does not preclude an individual element to have a large contribution $(p_i - q_i)^2$, which could cause the statistic to be too large and the tester to incorrectly reject. To avoid this, the term $\sqrt{mn} \cdot |p_i - q_i|$ in the denominator will "kick in" for any such element i, and make the ratio Z_i/f_i behave proportionally to $|p_i - q_i|/\sqrt{mn}$ instead of $(p_i - q_i)^2$, ensuring that the algorithm does not mistakenly reject "close" distributions due to any single large element contribution.

Remark 4 *For the identity testing problem, where the reference distribution q is known, we use the now-standard "splitting operation" of Diakonikolas and Kane (2016) (see Section A for details) to obtain distributions p' and q' over a domain of size $2n$ such that $\|p' - q'\|_1 = \|p - q\|_1$ and $\|q'\|_2 \leq 1/\sqrt{n}$. Moreover, samples from p and q can be used to simulate the same number of samples from distributions p' and q', respectively. We apply our tester on the modified distributions p' and q', instead of using it for p and q directly. As the new reference distribution q' is over a domain of size $2n$ and satisfies $\|q'\|_2 \leq \sqrt{2}/\sqrt{2n}$, this transformation lets us assume without loss of generality that the reference distribution q over $[n]$ in the identity testing problem is such that $\|q\|_2 \leq \sqrt{2/n}$.*

We now formally state the performance of Algorithm 1 for tolerant identity and equivalence testing, i.e., that it achieves near-optimal sample complexity in both cases.

Theorem 5 (Identity testing) *Let q be a known reference distribution and p be an unknown distributions, both over $[n]$. There exists an absolute constant $c > 0$ such that, for any $0 \leq \varepsilon_2 \leq 1$ and*

$0 \leq \varepsilon_1 \leq c\varepsilon_2$, *given*

$$\mathcal{O}\left(n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{\sqrt{n}}{\varepsilon_2^2}\right)$$

samples from each of p and q Algorithm 1 (after the splitting operation of Remark 4) distinguishes between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$ with probability at least $4/5$.

Theorem 6 (Equivalence testing) *Let p and q be two unknown distributions over $[n]$. There exists an absolute constant $c > 0$ such that, for any $0 \leq \varepsilon_2 \leq 1$ and $0 \leq \varepsilon_1 \leq c\varepsilon_2$, given*

$$\mathcal{O}\left(n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{\sqrt{n}}{\varepsilon_2^2} + \frac{n^{2/3}}{\varepsilon_2^{4/3}}\right)$$

samples from each of p and q Algorithm 1 distinguishes between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$ with probability at least $4/5$.

Note that for a unified exposition, we assumed in Theorem 5 that the algorithm is provided with samples even from the explicitly known reference distribution q. This is not a restriction, as given this explicit knowledge it is possible to efficiently sample from the distribution q.

We prove both the theorems in Appendix B.1.

3. Lower Bounds for Tolerant Testing

In this section, we derive our lower bounds on the "price of tolerance," i.e., on the increase in the sample complexity as a function of the parameters $\varepsilon_1, \varepsilon_2$. The main technical result is a lower bound for tolerant uniformity testing, from which the results for identity and equivalence will follow. In particular, we show:

Theorem 7 (The price of tolerance for uniformity testing) *For any n and $\varepsilon_1 < \varepsilon_2 < c$, for some universal constant $c > 0$, any tester which for any unknown distribution p over $[n]$ distinguishes between $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$ with probability at least $4/5$ must use*

$$\Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right)$$

samples from p.

By combining the above lower bound with previously known lower bounds for non-tolerant uniformity/identity testing (Paninski, 2008), we obtain:

Corollary 8 (Tolerant uniformity testing lower bound) *For any n and $0 \leq \varepsilon_1 < \varepsilon_2 < c$, for some universal constant $c > 0$, any tester which for any unknown distribution p over $[n]$ distinguishes between $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ vs $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$ with probability $\geq 4/5$ needs at least*

$$\Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \frac{\sqrt{n}}{\varepsilon_2^2}\right)$$

samples from p.

Similarly, by combining our lower bound with previously known lower bounds for non-tolerant equivalence testing (Valiant, 2011; Chan et al., 2014), we obtain:

Corollary 9 (Tolerant equivalence testing lower bound) *For any n and $0 \leq \varepsilon_1 < \varepsilon_2 < c$, for some universal constant $c > 0$, any tester which for any unknown distributions p and q, both over $[n]$, distinguishes between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$ with probability at least $4/5$ must use*

$$\Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)+\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2+\frac{\sqrt{n}}{\varepsilon_2^2}+\frac{n^{2/3}}{\varepsilon_2^{4/3}}\right)$$

samples.

3.1. The moment matching technique

The starting point for our proof of Theorem 7 is the moment matching technique first used in (Wu and Yang, 2016), which we considerably extend to address the "asymmetric" setting we must consider. We briefly review this technique here; we encourage the reader to refer to Appendix C for the details of our lower bound, and the extension to the asymmetric case we must handle. The first step is to consider the Poissonized version of the problem. Namely, given $\Theta(m)$ samples from a distribution $p = (p_1, \ldots, p_n)$, then with high probability, we can simulate a set of $\mathrm{Poi}(m)$ samples from the same distribution. Thus, without loss of generality, we may assume that we are given $\mathrm{Poi}(m)$ samples from p, and our goal is to distinguish with high probability given these samples whether $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ or $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$. A classical fact is that the result of sampling $\mathrm{Poi}(m)$ samples from p is identical in distribution to a draw from $(X_1, \ldots, X_n)$, where now the $X_i \sim \mathrm{Poi}(mp_i)$ are independent.

The high-level idea of the moment matching technique to construct two priors $\mathcal{U}$,$\mathcal{U}'$ over distributions on n elements so that with high probability two conditions hold. First, if $p \sim \mathcal{U}$ and $p' \sim \mathcal{U}'$, then with high probability, $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p' - \mathrm{Unif}_n\|_1 > \varepsilon_2$. Second, the result of (i) sampling a distribution $p \sim \mathcal{U}$ then (ii) sampling $\mathrm{Poi}(m)$ elements from p is close in TV distance to applying the same process to $\mathcal{U}'$. Specifically, the priors we construct will be product distributions, that is, $\mathcal{U} = \mathcal{P}^n$ and $\mathcal{U}' = (\mathcal{P}')^n$ for some positive univariate distributions $\mathcal{P}, \mathcal{P}'$. Then, ignoring some technical issues which we will address momentarily, the problem becomes: find distributions $\mathcal{P}, \mathcal{P}'$ supported on nonnegative values such that (1) $n\,\mathbb{E}_{\mathcal{P}}\,|p_i - 1/n| \leq \varepsilon_1$ and $n\,\mathbb{E}_{\mathcal{P}'}\,|p'_i - 1/n| > \varepsilon_2$, and (2) the following distance is small:

$$\mathrm{TV}\left(\mathop{\mathbb{E}}_{\mathcal{P}^n}\left(\mathrm{Poi}(mp_1), \ldots, \mathrm{Poi}(mp_n)\right),\ \mathop{\mathbb{E}}_{(\mathcal{P}')^n}\left(\mathrm{Poi}(mp'_1), \ldots, \mathrm{Poi}(mp'_n)\right)\right) = o(1)\ .$$

Note that, in view of the subaditivity of TV distance, this condition can be relaxed to the condition

$$\mathrm{TV}\left(\mathop{\mathbb{E}}_{\mathcal{P}}\mathrm{Poi}(mp_i),\ \mathop{\mathbb{E}}_{\mathcal{P}'}\mathrm{Poi}(mp'_i)\right) = o(1/n)\ . \tag{2}$$

While this will make later calculations much simpler, this introduces a couple of minor complications here. First, the vectors in the domain of $\mathcal{U}$,$\mathcal{U}'$ may not sum to 1, that is, $\mathcal{U}$ and $\mathcal{U}'$ may not actually be priors over *bona fide* distributions. However, if we additionally enforce that $\mathbb{E}_{V\sim\mathcal{P}}[V] = \mathbb{E}_{V'\sim\mathcal{P}'}[V'] = 1/n$, then under some mild conditions on the $\mathcal{P}, \mathcal{P}'$, by standard concentration arguments, the resulting vectors are very close to summing to 1 and thus form "approximate" distributions. One can then show that by slightly changing the construction, we can create priors over distributions that satisfy the desired properties. Second, the vectors p, p' in the domain of $\mathcal{U}, \mathcal{U}'$ may not deterministically satisfy the properties that $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p' - \mathrm{Unif}_n\|_1 > \varepsilon_2$. However, again by standard concentration inequalities, with high probability these random variables will not exceed their expectation by too much, and thus will satisfy these same constraints with high probability, perhaps relaxed by constant factors. We make this discussion more precise in the following theorem, whose (rather technical) proof is deferred to Section C.4:

Theorem 10 *Let $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$, and let n, m be positive integers and $m \geq c$, where $c > 0$ is an absolute constant. Suppose there exist random variables U, U' supported on the domain $[a, b]$ so that $b - a \leq \frac{\varepsilon_2^2}{1000}$, $\mathbb{E}[U] = \mathbb{E}[U'] = 1/n$, and*

$$\mathbb{E}\left[\left|U - \frac{1}{n}\right|\right] \leq \frac{\varepsilon_1}{n}\,, \qquad \text{and} \qquad \mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right] \geq \frac{\varepsilon_2}{n}\,. \tag{3}$$

Moreover, assume

$$\mathrm{TV}\big(\mathbb{E}\,\mathrm{Poi}(mU),\, \mathbb{E}\,\mathrm{Poi}(mU')\big) \leq \frac{1}{20n}\,. \tag{4}$$

Then, any tester which for any unknown distribution p distinguishes between $\|p - \mathrm{Unif}_n\|_1 \leq 25\varepsilon_1$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2/2$ with probability at least $4/5$ requires at least $m/2$ samples from p.

Thus, for given m, n and ε_1 the problem reduces to finding the maximum value of ε_2 for which we can construct a pair of random variables U and U' for which the assumptions of Theorem 10 hold. The next key insight is that we can further reduce the condition in (4) to designing two random variables with matching moments:

Lemma 11 ((Jiao et al., 2018, Lemma 32); see also (Wu and Yang, 2016)) *For any $\kappa \geq M \geq 0$, let Y, Y' be two random variables over $[\kappa - M, \kappa + M]$ so that $\mathbb{E}\,Y^i = \mathbb{E}\,Y'^i$ for all $i = 1, \dots, L$. Then, we have*

$$\mathrm{TV}\big(\mathbb{E}\,\mathrm{Poi}(Y), \mathbb{E}\,\mathrm{Poi}(Y')\big) \leq 2\left(\frac{eM}{\sqrt{\kappa(L+1)}}\right)^{L+1}.$$

With this lemma in place, our goal can be restated as follows: maximize $n \cdot \mathbb{E}\,|U_i' - 1/n|$ such that $n \cdot \mathbb{E}_{\mathcal{P}}\,|U - 1/n| \leq \varepsilon_1$ and the first L moments of U and U' match, where the support of U and U' is over $[\frac{\kappa - M}{m}, \frac{\kappa + M}{m}]$ for some $\kappa \geq M \geq 0$. The value of this maximum is a function of the parameters κ, M and L, whose values we choose later appropriately so that this function is maximized, while

$$2\left(\frac{eM}{\sqrt{\kappa(L+1)}}\right)^{L+1} \leq \frac{1}{20n} \tag{5}$$

holds, so that (4) is satisfied. We formulate the problem of maximizing $n \cdot \mathbb{E}\,|U' - 1/n|$ for any given choice of parameters as the following linear program over infinitely many variables, where we have used random variables V and V' to denote $n \cdot U$ and $n \cdot U'$, respectively,

$$\begin{aligned} \max\ \mathbb{E}\,|V' - 1| \text{ s.t. } \ & \mathbb{E}\,|V - 1| \leq \varepsilon_1 \text{ and} \\ & \mathbb{E}\,V = \mathbb{E}\,V' = 1, \text{ and} \\ & \mathbb{E}\,V^i = \mathbb{E}\,V'^i, i = 2, \dots, L, \text{ and} \\ & V, V' \in \left[\frac{n(\kappa - M)}{m}, \frac{n(\kappa + M)}{m}\right]. \end{aligned} \tag{6}$$

Let $\mathcal{L}(\varepsilon_1, n, m, M, \kappa, L)$ denote the value of the optimal solution of the above optimization problem. Observe that we do not need to find the exact solution to the above linear program: instead, any reasonable lower bound on the solution of the above optimization problem suffices. The next theorem gives one such lower bound. To state the theorem, we define

$$A := \frac{n(M+\kappa)}{m} - 1 - \varepsilon_1, \text{ and } B := \frac{n(M-\kappa)}{m} + 1 - \varepsilon_1. \tag{7}$$

Theorem 12 *For any κ, M, n, m, L, and ε_1, if for A, B, defined in (7), $0 < \varepsilon_1 \leq \min\left\{\frac{B}{4}, \frac{A}{4}\right\}$, then the value of optimal solution of (6) is lower bounded by*

$$\mathcal{L}(\varepsilon_1, n, m, M, \kappa, L) \geq \frac{1}{12} \max\left\{\sqrt{\varepsilon_1 \cdot \frac{A+B}{32L^2}}, \sqrt{\varepsilon_1 \cdot \frac{\sqrt{AB}}{16L}}\right\}. \tag{8}$$

To prove this theorem, which is the key technical component in the proof of lower bound, we had to completely diverge from the previous works, which bound the solution of the dual that arise for their applications by finding the best uniform approximation of a given function by a low-degree polynomial (for example in (Wu and Yang, 2016)). Indeed, the asymmetry in the range of V and V' around zero makes it difficult to find such a low-degree polynomial. Instead we prove the theorem in Appendix C.2 from first principles by a novel use of classic tools from polynomial approximation theory. In Appendix C.1, we then use this theorem to prove the distribution testing lower bounds.

Acknowledgments

The authors would like to thank Sivaraman Balakrishnan for helpful pointers on the Statistics literature and the anonymous reviewers for helpful comments. Gautam kamath is supported by an NSERC Discovery Grant, an unrestricted gift from Google, and a University of Waterloo startup grant.

References

Alberto Abadie. Statistical nonsignificance in empirical economics. *American Economic Review: Insights*, 2(2):193–208, 2020.

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ISIT '14, pages 3200–3204, Washington, DC, USA, 2014. IEEE Computer Society.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 11–21. JMLR, Inc., 2017.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society.

Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4:1–4:25, 2013.

James O. Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

Serge Bernstein. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, volume 4. Hayez, imprimeur des académies royales, 1912.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The american mathematical monthly*, 107(4):353–357, 2000.

Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2611–2619. Curran Associates, Inc., 2015.

Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Proceedings of the 32nd Computational Complexity Conference*, CCC '17, pages 28:1–28:40, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, (9):1–100, 2020a.

Clément L. Canonne. A short note on learning discrete distributions. *CoRR*, abs/2002.11457, 2020b.

Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Exploring the gap between tolerant and non-tolerant distribution testing. *CoRR*, abs/2110.09972, 2021.

Siu On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2747–2764, Philadelphia, PA, USA, 2018. SIAM.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, ICALP '18, pages 41:1–41:14, 2018.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chicago Journal of Theoretical Computer Science*, 1:1–21, 2019.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53nd Annual ACM Symposium on the Theory of Computing*, STOC '21, New York, NY, USA, 2021. ACM.

Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23(15), 2016.

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

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. In *Proceedings of the 2016 International Symposium on Information Theory and Its Applications*, ISITA '16, pages 256–260, Washington, DC, USA, 2016. IEEE Computer Society.

Yuri Izmailovich Ingster. Minimax detection of a signal in ℓ_p metrics. *Journal of Mathematical Sciences*, 68(4):503–515, 1994.

Yuri Izmailovich Ingster. On testing a hypothesis which is close to a simple hypothesis. *Teoriya Veroyatnostei i ee Primeneniya*, 45:356–368, 2000.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2017.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

Gautam Kamath. *Modern Challenges in Distribution Testing*. PhD thesis, Massachusetts Institute of Technology, September 2018.

Sudeep Kamath, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 1066–1100, 2015.

Wladimir Markoff. Ober polynome, die in einem gegebenen intervalle moglichst wenig von null abweichen. *Ann., 77, 213-258 (1892)(translation and condenstation by J. Grossman of Russian article published*, 1892.

Alon Orlitsky, Ananda Theerta Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006.

Calyampudi Radhakrishna Rao and Miodrag M Lovric. Testing point null hypothesis of a normal mean and the truth: 21st century perspective. *Journal of Modern Applied Statistical Methods*, 15 (2):2–21, 2016.

Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

R Tyrrell Rockafellar. *Conjugate Duality and Optimization*. SIAM, 1974.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(179), 2010a.

Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(180), 2010b.

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 *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011a. ACM.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '11, pages 403–412, Washington, DC, USA, 2011b. IEEE Computer Society.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6): 1927–1968, 2011.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 2018.

Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, 17(4):402–586, 2020.

Appendix A. Details on the "splitting" operation

Given an explicit reference distribution q over $[n]$, we describe the splitting operation with respect to q, as introduced in Diakonikolas and Kane (2016). For $i \in [n]$, let $a_i \coloneqq 1 + \lfloor nq_i \rfloor$ and $D \coloneqq \{(i,j) : i \in [n], j \in [a_i]\}$. The splitting operation with respect to q maps any given distribution p over $[n]$ to a new distribution $p^{S(q)}$ over the new domain D such that the new distribution $p^{S(q)}$ assigns the probability $\frac{p_i}{a_i}$ to element (i,j).

We note a few properties of the splitting operation:

1. The new domain is at most twice as large: indeed, $|D| = \sum_{i=1}^{n}(1 + \lfloor nq_i \rfloor) \leq \sum_{i=1}^{n}(nq_i + 1) = 2n$.

2. If q is known, then m i.i.d. samples from an unknown distribution p can be used to simulate the m i.i.d. from $p^{S(q)}$, by (independently for each) mapping a sample $i \in [n]$ to a (i,j), for j chosen uniformly at random in $[a_i]$.

3. The resulting distribution obtained by applying splitting operation w.r.t. q on itself has small ℓ_2 norm,
$$\|q^{S(q)}\|_2^2 = \sum_{(i,j)\in D} (q_{i,j}^{S(q)})^2 = \sum_{i=1}^{n} \sum_{j\in[a_i]} \left(\frac{q_i}{a_i}\right)^2 = \sum_{i=1}^{n} \frac{q_i^2}{a_i} \leq \sum_{i=1}^{n} \frac{q_i}{n} = \frac{1}{n} \leq \frac{2}{|D|}\,.$$

4. The pairwise ℓ_1 (and thus total variation) distances between any two distributions p and p' are preserved after the splitting operation, namely for any distributions p, p' over $[n]$,
$$\|p - p'\|_1 = \|p^{S(q)} - p'^{S(q)}\|_1\,;$$
this follows from observing that
$$\|p^{S(q)} - p'^{S(q)}\|_1 = \sum_{i=1}^{n} \sum_{j\in[a_i]} |p_{i,j}^{S(q)} - p_{i,j}'^{S(q)}| = \sum_{i=1}^{n} \sum_{j\in[a_i]} \left|\frac{p(i)}{a_i} - \frac{p'(i)}{a_i}\right| = \sum_{i=1}^{n} |p_i - p_i'|.$$

Appendix B. Upper bound proofs

B.1. Analysis of Algorithm 1

This section is devoted to the proofs of Theorems 5 and 6, which are both established in a similar manner.

Observe that, following the Poissonization, all Z_i's and $\widehat{f}_i$'s are independent random variables. From the properties of Poisson distribution, it is not hard to check that the expectation and variance of the Z_i's are given by
$$\mathbb{E}[Z_i] = m^2|p_i - q_i|^2, \tag{9}$$
$$\mathrm{Var}[Z_i] = 4m^3(p_i - q_i)^2(p_i + q_i) + 2m^2(p_i + q_i)^2. \tag{10}$$

Next, using the independence of Z_i and $\widehat{f}_i$'s, we get that the conditional expectation of Z is
$$\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i \in [n]\Big] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{Z_i}{\widehat{f}_i} \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right] = \sum_{i=1}^{n} \frac{\mathbb{E}[Z_i]}{\widehat{f}_i}, \tag{11}$$

while its conditional variance is given by

$$\operatorname{Var}\left[Z \middle| \widehat{f}_i \text{ for } i \in [n]\right] = \operatorname{Var}\left[\sum_{i=1}^{n} \frac{Z_i}{\widehat{f}_i} \middle| \widehat{f}_i \text{ for } i \in [n]\right] = \sum_{i=1}^{n} \frac{\operatorname{Var}(Z_i)}{\widehat{f}_i^2}. \tag{12}$$

To prove the optimality of the tester, we first bound the conditional expectation and variance of Z. These bounds differ for the regimes $m \geq n$ and $m \leq n$, and are characterized in Lemmas 13 and 14, respectively.

Lemma 13 *There exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. For $m \geq n$, and any distributions p and q over $[n]$, the following bounds simultanously hold with probability at least $9/10$:*

$$c_1 \min\left(\frac{m^{3/2}\|p-q\|_1}{n^{1/2}}, \frac{m^2\|p-q\|_1{}^2}{n}\right) \leq \mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \textit{ for } i \in [n]\right] \leq c_2 \frac{m^{3/2}\|p-q\|_1}{n^{1/2}},$$

and $\operatorname{Var}\left[Z \;\middle|\; \widehat{f}_i \textit{ for } i \in [n]\right] \leq c_3 \frac{m^2}{n}$.

Lemma 14 *There exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. For $m \leq n$, and any distributions p and q over $[n]$, the following bounds simultanously hold with probability at least $9/10$:*

$$c_1 \frac{m^2\|p-q\|_1{}^2}{n} \leq \mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \textit{ for } i \in [n]\right] \leq c_2 m \|p-q\|_1,$$

and $\operatorname{Var}\left[Z \;\middle|\; \widehat{f}_i \textit{ for } i \in [n]\right] \leq c_3 m$. *Additionally,*

$$\operatorname{Var}\left[Z \;\middle|\; \widehat{f}_i \textit{ for } i \in [n]\right] \leq \frac{1}{40}(\mathbb{E}[Z \mid \widehat{f}_i \textit{ for } i \in [n]])^2 + 324\,\mathbb{E}[Z \mid \widehat{f}_i \textit{ for } i \in [n]] + 648 m^2 \|q\|_2^2.$$

We prove these lemmata in Section B.2. We now show that, assuming these statements, we can establish Theorems 5 and 6. We handle the cases $m \geq n$ and $m < n$ separately.

Proof of the theorems for $m \geq n$: Using Lemma 13, we show that for any $m \geq n$ such that $m = \Omega\left(n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \frac{\sqrt{n}}{\varepsilon_2^2}\right)$ the estimator correctly distinguishes between $\|p-q\| \leq \varepsilon_1$ vs $\|p-q\| \geq \varepsilon_2$ with probability at least $8/10$.

Applying Chebyshev's inequality to the conditional expectation and variance and using Lemma 13, we get that, with probability $\geq 8/10$,

$$Z \geq c_1 \min\left(\frac{m^{3/2}\|p-q\|_1}{n^{1/2}}, \frac{m^2\|p-q\|_1{}^2}{n}\right) - \sqrt{10c_3}\frac{m}{\sqrt{n}}, \tag{13}$$

and

$$Z \leq c_2 \frac{m^{3/2}\|p-q\|_1}{n^{1/2}} + \sqrt{10c_3}\frac{m}{\sqrt{n}}. \tag{14}$$

- In the case $\|p-q\|_1 \geq \varepsilon_2$, the lower bound in Equation (13) reduces to

$$Z \geq c_1 \min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right) - \sqrt{10c_3}\frac{m}{\sqrt{n}} \geq \frac{c_1}{2}\min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right),$$

  the last step as long as $m \geq C\sqrt{n}/\varepsilon_2^2$ for $C := \max(2\sqrt{10c_3}/c_1, 40c_3/c_1^2)$. Therefore, with probability at least $8/10$ the tester correctly outputs that $\|p-q\|_1 \geq \varepsilon_2$.

- In the case, $\|p-q\|_1 \leq \varepsilon_1$, the upper bound in Equation (14) reduces to

$$Z \leq c_2\frac{m^{3/2}\varepsilon_1}{n^{1/2}} + \sqrt{10c_3}\frac{m}{\sqrt{n}} \leq \frac{c_1}{4}\min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$$

  where we used that $m \geq C'\sqrt{n}/\varepsilon_2^2$ for $C := \max(\frac{8\sqrt{10c_3}}{c_1}, \frac{640c_3}{c_1^2})$ to ensure that $\sqrt{10c_3}\frac{m}{\sqrt{n}} \leq \frac{c_1}{8}\min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$, and that (i) $\varepsilon_1 \leq \frac{c_1}{8c_2}\varepsilon_2$ and (ii) $m \geq C''\cdot n\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2$ with $C' = 64c_2^2/c_1^2$ to ensure that $c_2\frac{m^{3/2}\varepsilon_1}{n^{1/2}} \leq \frac{c_1}{8}\min\left(\frac{m^{3/2}\varepsilon_2}{n^{1/2}}, \frac{m^2\varepsilon_2^2}{n}\right)$. Therefore, with probability at least $8/10$ the tester correctly outputs that $\|p-q\|_1 \leq \varepsilon_1$.

This proves the two theorems for the case $m \geq n$. We next turn to the case $m \leq n$.

Proof of the theorems for $m < n$: The argument for this case is similar to the previous, using Lemma 14 instead of Lemma 13. We show that for any $m \leq n$ such that $m = \Omega\left(n\frac{\varepsilon_1}{\varepsilon_2^2} + \min\left\{\frac{n\|q\|_2}{\varepsilon_2^2}, \frac{n^{2/3}}{\varepsilon_2^{4/3}}\right\}\right)$ the estimator correctly distinguishes between $\|p-q\| \leq \varepsilon_1$ and $\|p-q\| \geq \varepsilon_2$ with probability at least $8/10$. This in turn follows from computations nearly identical to the ones above, which we thus omit in the interest of space.

The proofs for the two cases, combined with the fact that for identity testing we can as discussed before assume without loss of generality that $\|q\|_2 \leq \sqrt{2/n}$, establish Theorems 5 and 6. □

B.2. Proof of Lemmas 13 and 14

In this section, we give the proof of the remaining two pieces in our analysis of Algorithm 1, Lemmas 13 and 14. The following lemma will be useful to lower bound the conditional expectation $\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right]$.

Lemma 15 *There exist absolute constants $c_1, c_2, c_3 > 0$ such that, for every m, n, and $i \in [n]$,*

$$\mathbb{E}[\widehat{f}_i] \leq c_1 f_i, \qquad \mathbb{E}[\widehat{f}_i^{-1}] \leq \frac{c_2}{f_i}, \quad \textit{and } \mathbb{E}[\widehat{f}_i^{-2}] \leq \frac{c_3}{f_i^2}.$$

Proof We use the following two concentration bounds, which provide exponential tail bounds on our estimates $\widehat{f}_i$ of the of f_i's. The proofs of those two claims are quite technical, and rely on a careful case distinction along with standard concentration properties of Poisson random variables. We provide them in Section B.3.

Lemma 16 *There exists $c > 0$ such that, for every m, n, $t > 3$, and $i \in [n]$, $\Pr[\widehat{f}_i > tf_i] \leq e^{-ct}$.*

Lemma 17 *There exists $c' > 0$ such that, for every m, n, $t > 2$, and $i \in [n]$,* $\Pr[\widehat{f}_i < \frac{f_i}{t}] \leq e^{-c't}$.

Given the above two results, the proof is straightforward: indeed, for every $i \in [n]$ we have, using Lemma 16,

$$\mathbb{E}[\widehat{f}_i] = \int_0^\infty \Pr[\widehat{f}_i > u]du = f_i \int_0^\infty \Pr[\widehat{f}_i > tf_i]dt \leq f_i\left(\int_0^3 dt + \int_3^\infty e^{-ct}dt\right) = f_i\left(3 + \frac{e^{-3c}}{c}\right)$$

while, from Lemma 17,

$$\mathbb{E}[\widehat{f}_i^{-1}] = \int_0^\infty \Pr[\widehat{f}_i^{-1} > u]du = \frac{1}{f_i}\int_0^\infty \Pr[\widehat{f}_i < f_i/t]dt \leq \frac{1}{f_i}\left(2 + \int_2^\infty e^{-c't}dt\right) = \frac{1}{f_i}\left(2 + \frac{e^{-2c'}}{c'}\right)$$

and, similarly,

$$\mathbb{E}[\widehat{f}_i^{-2}] = \int_0^\infty \Pr[\widehat{f}_i^{-2} > u]du = \frac{2}{f_i^2}\int_0^\infty \Pr[\widehat{f}_i < f_i/t]t\,dt \leq \frac{2}{f_i^2}\left(2 + \int_2^\infty te^{-c't}dt\right) = \frac{2}{f_i^2}\left(2 + \frac{(2c'+1)e^{-2c'}}{c'^2}\right)$$

which, given that c, c' are just positive constants, is what we set out to prove. ■

We also require the following simple inequality.

Fact 18 *For any real $(a_i)_{i=1}^n$ and positive $(b_i)_{i=1}^n$,*

$$\sum_{i=1}^n \frac{a_i^2}{b_i} \geq \frac{(\sum_{i=1}^n |a_i|)^2}{\sum_{i=1}^n b_i}.$$

Proof The result follows from applying Cauchy–Schwarz to $\sum_{i=1}^n |a_i| = \sum_{i=1}^n \sqrt{b_i}|a_i|/\sqrt{b_i}$. ■

From the above fact and (9), it follows that

$$\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right] = \sum_{i=1}^n \frac{\mathbb{E}[Z_i]}{\widehat{f}_i} = \sum_{i=1}^n \frac{m^2(p_i - q_i)^2}{\widehat{f}_i} \geq \frac{m^2(\sum_{i=1}^n |p_i - q_i|)^2}{\sum_{i=1}^n \widehat{f}_i} = \frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n \widehat{f}_i}.$$

Moreover, by definition the random variables $\widehat{f}_i$ are non-negative, and thus, applying the Markov inequality we get that

$$\sum_{i=1}^n \widehat{f}_i \leq 30\sum_{i=1}^n \mathbb{E}[\widehat{f}_i]$$

with probability at least $1 - 1/30$. Combined with Lemma 15, this means that, with probability at least $1 - 1/30$,

$$\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right] \geq \frac{m^2\|p-q\|_1^2}{30c_1\sum_{i=1}^n f_i}. \tag{15}$$

Next, applying the Markov's inequality for the non-negative random variable $\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right]$, we get that, with probability at least $1 - 1/30$,

$$\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right] \leq 30\,\mathbb{E}\left[\mathbb{E}\left[Z \;\middle|\; \widehat{f}_i \text{ for } i \in [n]\right]\right] = 30\,\mathbb{E}\left[\sum_{i=1}^n \frac{\mathbb{E}[Z_i]}{\widehat{f}_i}\right] \leq 30c_2\sum_{i=1}^n \frac{m^2(p_i - q_i)^2}{f_i}. \tag{16}$$

Finally, considering the non-negative random variable $\operatorname{Var}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i \in [n]\Big]$, we again get that, with probability at least $1 - 1/30$,

$$\operatorname{Var}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i \in [n]\Big] \le 30\,\mathbb{E}\Big[\operatorname{Var}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i \in [n]\Big]\Big] = 30\,\mathbb{E}\left[\sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{\widehat{f}_i^2}\right] \le 30c_3 \sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{f_i^2}. \tag{17}$$

By a union bound, we get that the guarantees of (15), (16), and (17) simultaneously hold with probability at least $1 - 3 \cdot \frac{1}{30} = \frac{9}{10}$. Importantly, the RHS in all three bounds only depend on the deterministic quantities f_i's, instead of the random variables $\widehat{f}_i$'s. We bound each of these RHS in the next two lemmas, for $m \ge n$ and $m \le n$, respectively.

Lemma 19 *For any $m \ge n$ and distributions p and q over $[n]$ the following holds: (1) $\sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{f_i^2} \le \frac{10m^2}{n}$, (2) $\sum_{i=1}^n \frac{m^2(p_i-q_i)^2}{f_i} \le \frac{m^{3/2}\|p-q\|_1}{n^{1/2}}$, and (3) $\frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n f_i} \ge \min\left(\frac{m^{3/2}\|p-q\|_1}{2n^{1/2}}, \frac{m^2\|p-q\|_1{}^2}{6n}\right)$.*

Proof First, we upper bound $\sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{f_i^2}$: from (10), we get

$$\begin{aligned}
\sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{f_i^2} &= \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i) + 2m^2(p_i+q_i)^2}{f_i^2} \\
&= \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i) + 2m^2(p_i+q_i)^2}{\Big(\max\{\sqrt{mn}\cdot|p_i-q_i|, n\cdot(p_i+q_i), 1\}\Big)^2} \\
&\le \sum_{i=1}^n \frac{4m^3(p_i+q_i)}{mn} + \sum_{i=1}^n \frac{2m^2}{n^2} = \frac{8m^2}{n} + \frac{2m^2}{n} = \frac{10m^2}{n}.
\end{aligned}$$

Next, we prove the second inequality:

$$\begin{aligned}
\sum_{i=1}^n \frac{m^2(p_i-q_i)^2}{f_i} &= \sum_{i=1}^n \frac{m^2|p_i-q_i|^2}{\max\{\sqrt{mn}\cdot|p_i-q_i|, n\cdot(p_i+q_i), 1\}} \\
&\le \sum_{i=1}^n \frac{m^{3/2}|p_i-q_i|}{n^{1/2}} = \frac{m^{3/2}\|p-q\|_1}{n^{1/2}}.
\end{aligned}$$

Finally, we prove the last inequality:

$$\begin{aligned}
\frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n f_i} &= \frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n \max\{\sqrt{mn}\cdot|p_i-q_i|, n\cdot(p_i+q_i), 1\}} \\
&\ge \frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n(\sqrt{mn}\cdot|p_i-q_i| + n\cdot(p_i+q_i) + 1)} \\
&= \frac{m^2\|p-q\|_1^2}{\sqrt{mn}\cdot\|p-q\|_1 + 2n + n} \\
&\ge \min\left(\frac{m^{3/2}\|p-q\|_1}{2n^{1/2}}, \frac{m^2\|p-q\|_1{}^2}{6n}\right).
\end{aligned}$$

■

Lemma 20 *For any $m \leq n$ and distributions p and q over $[n]$ the following holds: $\sum_{i=1}^n \frac{\mathrm{Var}(Z_i)}{f_i^2} \leq 24m$, (2) $\sum_{i=1}^n \frac{m^2(p_i-q_i)^2}{f_i} \leq m\|p-q\|_1$, and (3) $\frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n f_i} \geq \frac{m^2\|p-q\|_1^2}{3n}$.*

Proof As before, we first upper bound $\sum_{i=1}^n \frac{\mathrm{Var}(Z_i)}{f_i^2}$:

$$\begin{aligned}
\sum_{i=1}^n \frac{\mathrm{Var}(Z_i)}{f_i^2} &= \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i)+2m^2(p_i+q_i)^2}{f_i^2} \\
&= \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i)+2m^2(p_i+q_i)^2}{\Big(\max\{m\cdot(p_i+q_i),1\}\Big)^2} \\
&\leq \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i)+4m^2(p_i-q_i)^2+8m^2q_i^2}{\max\{m^2\cdot(p_i+q_i)^2,1\}} \\
&\qquad \text{(as } (a+b)^2 \leq 2(a-b)^2+4b^2\text{)} \\
&\leq \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i)}{m^2\cdot(p_i+q_i)^2} + \sum_{i=1}^n \frac{4m^2(p_i-q_i)^2}{m\cdot(p_i+q_i)} + \sum_{i=1}^n \frac{8m^2q_i^2}{\max\{m^2(p_i+q_i)^2,1\}} \\
&\leq 4m\sum_{i=1}^n |p_i-q_i| + 4m\sum_{i=1}^n |p_i-q_i| + \sum_{i=1}^n \frac{8m^2q_i^2}{\max\{m(p_i+q_i),1\}} \\
&\qquad (\max\{x^2,1\} \geq \max\{x,1\}) \\
&\leq 8m\|p-q\|_1 + \sum_{i=1}^n 8mq_i \\
&\leq 16m + 8m = 24m\,.
\end{aligned}$$

Next, we prove the second inequality:

$$\sum_{i=1}^n \frac{m^2(p_i-q_i)^2}{f_i} = \sum_{i=1}^n \frac{m^2|p_i-q_i|^2}{\max\{m\cdot(p_i+q_i),1\}} \leq \sum_{i=1}^n m|p_i-q_i| = m\|p-q\|_1.$$

Finally, we prove the last inequality:

$$\frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n f_i} = \frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n \max\{m\cdot(p_i+q_i),1\}} \geq \frac{m^2\|p-q\|_1^2}{\sum_{i=1}^n (m\cdot(p_i+q_i)+1)} = \frac{m^2\|p-q\|_1^2}{2m+n} \geq \frac{m^2\|p-q\|_1^2}{3n}.$$

■

It only remains to establish the last part of Lemma 14, which we do next.

$$\begin{aligned}
\operatorname{Var}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big] &= \sum_{i=1}^n \frac{\operatorname{Var}(Z_i)}{\widehat{f}_i^2} = \sum_{i=1}^n \frac{4m^3(p_i-q_i)^2(p_i+q_i)+2m^2(p_i+q_i)^2}{\widehat{f}_i^2}\\
&\overset{\text{(a)}}{\leq} 4m^3\left(\sum_{i=1}^n \frac{(p_i-q_i)^4}{\widehat{f}_i^2}\right)^{1/2}\left(\sum_{i=1}^n \frac{(p_i+q_i)^2}{\widehat{f}_i^2}\right)^{1/2} + \sum_{i=1}^n \frac{2m^2(p_i+q_i)^2}{\widehat{f}_i^2}\\
&\overset{\text{(b)}}{\leq} 4m^3\left(\sum_{i=1}^n \frac{(p_i-q_i)^2}{\widehat{f}_i}\right)\left(\sum_{i=1}^n \frac{(p_i+q_i)^2}{\widehat{f}_i^2}\right)^{1/2} + \sum_{i=1}^n \frac{2m^2(p_i+q_i)^2}{\widehat{f}_i^2}\\
&= 4\left(m^2\sum_{i=1}^n \frac{(p_i-q_i)^2}{\widehat{f}_i}\right)\left(m^2\sum_{i=1}^n \frac{(p_i+q_i)^2}{\widehat{f}_i^2}\right)^{1/2} + \sum_{i=1}^n \frac{2m^2(p_i+q_i)^2}{\widehat{f}_i^2}
\end{aligned}$$

where step (a) is the Cauchy–Schwarz inequality, and (b) is monotonicity of ℓ_p norms: for any vector u, $\|u\|_2 \leq \|u\|_1$. We can then continue as follows, making the expectation appear:

$$\begin{aligned}
\operatorname{Var}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big] &= 4\Big(\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big]\Big)\left(\sum_{i=1}^n \frac{m^2(p_i+q_i)^2}{\widehat{f}_i^2}\right)^{1/2} + \sum_{i=1}^n \frac{2m^2(p_i+q_i)^2}{\widehat{f}_i^2}\\
&\overset{\text{(c)}}{\leq} \frac{1}{40}\Big(\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big]\Big)^2 + (160+2)\sum_{i=1}^n \frac{m^2(p_i+q_i)^2}{\widehat{f}_i^2}\\
&\overset{\text{(d)}}{\leq} \frac{1}{40}\Big(\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big]\Big)^2 + 162\sum_{i=1}^n \frac{2m^2(p_i-q_i)^2}{\widehat{f}_i^2} + 162\sum_{i=1}^n \frac{4m^2q_i^2}{\widehat{f}_i^2}\\
&\overset{\text{(e)}}{\leq} \frac{1}{40}\Big(\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big]\Big)^2 + 324\sum_{i=1}^n \frac{m^2(p_i-q_i)^2}{\widehat{f}_i} + 648\sum_{i=1}^n m^2q_i^2\\
&= \frac{1}{40}\Big(\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big]\Big)^2 + 324\,\mathbb{E}\Big[Z \;\Big|\; \widehat{f}_i \text{ for } i\in[n]\Big] + 648m^2\|q\|_2^2,
\end{aligned}$$

where (c) uses $2ab \leq a^2+b^2$, (d) uses $(a+b)^2 \leq 2(a-b)^2+4b^2$, and finally (e) uses $\widehat{f}_i \geq 1$. □

B.3. Proofs of Lemmas 16 and 17

The following standard bound on the concentration of Poisson random variables will be useful:

Theorem 21 *Let $X \sim \operatorname{Poi}(\lambda)$ be a Poisson random variable for some $\lambda > 0$. Then for any $x > 0$,*

$$\Pr\left[|X-\lambda| \geq x\right] \leq \exp\Big(-\Omega\Big(\min\Big\{x, \frac{x^2}{\lambda}\Big\}\Big)\Big).$$

Next, we prove Lemma 16.

Proof of Lemma 16 First we prove the lemma for the simpler of the two cases when $m < n$ and then later for $m \geq n$.

Proof for the regime $m < n$:

$$
\begin{aligned}
\Pr\left[\widehat{f}_i > tf_i\right] &\overset{\text{(a)}}{=} \Pr\left[\max\{X_i + Y_i, 1\} > t\max\{m(p_i+q_i),1\}\right]\\
&= \Pr\left[X_i + Y_i > t\max\{m(p_i+q_i),1\}\right]\\
&= \Pr\left[X_i + Y_i - m(p_i+q_i) > t\max\{m(p_i+q_i),1\} - m(p_i+q_i)\right]\\
&= \Pr\left[X_i + Y_i - m(p_i+q_i) > (t-1)\max\{m(p_i+q_i),1\}\right]\\
&\overset{\text{(b)}}{\leq} \exp\left(-\Omega\left(\min\left\{(t-1)\max\{m(p_i+q_i),1\}, \frac{(t-1)^2\max\{m(p_i+q_i)^2,1\}}{\max\{m(p_i+q_i),1\}}\right\}\right)\right)\\
&\leq \exp\bigl(-\Omega\bigl(\min\bigl\{(t-1),(t-1)^2\bigr\}\bigr)\bigr)\\
&\overset{\text{(c)}}{=} \exp(-\Omega(t)),
\end{aligned}
$$

here (a) uses definition of f_i, (b) uses the fact that $\tilde{X} + \tilde{Y} \sim \mathrm{Poi}(mp_i + mq_i)$ and the Poisson concentration bound in Theorem 21, and (c) uses $t > 1$.

Next we prove the lemma for the other case when $m \geq n$.

Proof for the regime $m \geq n$: We first bound $\Pr\left[\widehat{f}_i > tf_i\right]$ by sum of three different terms, and then later we bound each term one by one.

$$
\begin{aligned}
&\Pr\left[\widehat{f}_i > tf_i\right]\\
&= \Pr\left[\max\left\{\frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}}, \frac{\tilde{X}_i + \tilde{Y}_i}{m/n}, 1\right\} > tf_i\right]\\
&\overset{\text{(a)}}{\leq} \Pr\left[\frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}} > tf_i\right] + \Pr\left[\frac{\tilde{X}_i + \tilde{Y}_i}{m/n} > tf_i\right]\\
&\overset{\text{(b)}}{\leq} \Pr\left[\frac{|\tilde{X}_i - mp_i| + |\tilde{Y}_i - mq_i| + |mp_i - mq_i|}{\sqrt{m/n}} > tf_i\right] + \Pr\left[\frac{\tilde{X}_i + \tilde{Y}_i}{m/n} > tf_i\right]\\
&= \Pr\left[|\tilde{X}_i - mp_i| + |\tilde{Y}_i - mq_i| > tf_i\sqrt{m/n} - m|p_i - q_i|\right] + \Pr\left[\tilde{X}_i + \tilde{Y}_i > tf_i\frac{m}{n}\right]\\
&\leq \Pr\left[\max\{|\tilde{X}_i - mp_i|, |\tilde{Y}_i - mq_i|\} > \frac{tf_i\sqrt{m/n} - m|p_i - q_i|}{2}\right] + \Pr\left[\tilde{X}_i + \tilde{Y}_i > tf_i\frac{m}{n}\right]\\
&\overset{\text{(c)}}{\leq} \Pr\left[|\tilde{X}_i - mp_i| > \frac{tf_i\sqrt{m/n} - m|p_i - q_i|}{2}\right] + \Pr\left[|\tilde{Y}_i - mq_i| > \frac{tf_i\sqrt{m/n} + -m|p_i - q_i|}{2}\right]\\
&+ \Pr\left[\tilde{X}_i + \tilde{Y}_i > tf_i\frac{m}{n}\right], \qquad (18)
\end{aligned}
$$

where inequalities (a) and (c) use union bound and (b) uses triangle inequality.

To obtain an upper bound we bound each term in the above equation. Next, we bound the first term.

$$\begin{aligned}
&\Pr\left[|\tilde{X}_i - mp_i| > \frac{t}{2} f_i \sqrt{m/n} - \frac{m|p_i - q_i|}{2}\right] \\
&\overset{(a)}{=} \Pr\left[|\tilde{X}_i - mp_i| > \frac{t}{2}\max\{m|p_i - q_i|, \sqrt{mn}(p_i + q_i), \sqrt{m/n}\} - \frac{m|p_i - q_i|}{2}\right] \\
&\overset{(b)}{\leq} \Pr\left[|\tilde{X}_i - mp_i| > \frac{tm|p_i - q_i| + t\sqrt{mn}(p_i + q_i) + t\sqrt{m/n}}{6} - \frac{m|p_i - q_i|}{2}\right] \\
&= \Pr\left[|\tilde{X}_i - mp_i| > \frac{(t-3)m|p_i - q_i| + t\sqrt{mn}(p_i + q_i) + t\sqrt{m/n}}{6}\right] \\
&\overset{(c)}{\leq} \Pr\left[|\tilde{X}_i - mp_i| > \frac{t\sqrt{mn}(p_i + q_i) + t\sqrt{m/n}}{6}\right] \\
&\overset{(d)}{\leq} 2\exp\left(-\Omega\left(\min\left\{\frac{t\sqrt{mn}(p_i + q_i) + t\sqrt{m/n}}{6}, \frac{(t\sqrt{mn}(p_i + q_i) + t\sqrt{m/n})^2}{36mp_i}\right\}\right)\right) \\
&\leq 2\exp\left(-\Omega\left(\min\left\{\frac{t\sqrt{m/n}}{6}, \frac{t^2 mn(p_i + q_i)^2 + t^2 \cdot (m/n)}{36mp_i}\right\}\right)\right) \\
&\leq 2\exp\left(-\Omega\left(\min\left\{\frac{t\sqrt{m/n}}{6}, \frac{t^2 np_i}{36} + \frac{t^2}{36np_i}\right\}\right)\right) \\
&\overset{(e)}{\leq} \exp(-\Omega(t)),
\end{aligned}$$

where (a) uses the definition of f_i, (b) follows from the fact that $\frac{a+b+c}{3} \leq \max\{a, b, c\}$, inequality (c) uses $t \geq 3$, inequality (d) uses the fact that $\tilde{X} \sim \text{Poi}(mp_i)$ and the Poisson concentration bound in Theorem 21, and finally (e) uses $m \geq n$ and the fact that $x + 1/x \geq 2$ for any $x \geq 0$.

Note that because of the symmetry the above bound will also apply on the second term, namely

$$Pr\left[|\tilde{X}_i - mq_i| > \frac{t}{2} f_i \sqrt{m/n} - \frac{m|p_i - q_i|}{2}\right] \leq \exp(-\Omega(t)).$$

Next we bound the last term in Equation (18) to complete the proof of the first concentration inequality.

$$
\begin{aligned}
&\Pr\left[\tilde{X}_i + \tilde{Y}_i > t f_i \frac{m}{n}\right] \\
&\overset{(a)}{=} \Pr\left[\tilde{X}_i + \tilde{Y}_i > t \max\{\frac{m^{3/2}}{n^{1/2}}|p_i - q_i|, m(p_i+q_i), \frac{m}{n}\}\right] \\
&\le \Pr\left[\tilde{X}_i + \tilde{Y}_i > t \max\{m(p_i+q_i), \frac{m}{n}\}\right] \\
&\overset{(b)}{\le} \Pr\left[\tilde{X}_i + \tilde{Y}_i - m(p_i+q_i) > \frac{tm(p_i+q_i) + t\frac{m}{n}}{2} - m(p_i+q_i)\right] \\
&= \Pr\left[\tilde{X}_i + \tilde{Y}_i - m(p_i+q_i) > \frac{(t-2)m(p_i+q_i) + t\frac{m}{n}}{2}\right] \\
&\overset{(c)}{\le} 2\exp\left(-\Omega\Big(\min\left\{\frac{(t-2)m(p_i+q_i) + t\frac{m}{n}}{2}, \frac{((t-2)m(p_i+q_i) + t\frac{m}{n})^2}{4m(p_i, q_i)}\right\}\Big)\right) \\
&\overset{(d)}{\le} 2\exp\left(-\Omega\Big(\min\left\{t\frac{m}{2n}, \frac{((t-2)m(p_i+q_i) + (t-2)\frac{m}{n})^2}{4m(p_i+q_i)}\right\}\Big)\right) \\
&\overset{(e)}{\le} 2\exp\left(-\Omega\Big(\min\left\{t\frac{m}{2n}, \frac{(t-2)^2 m(p_i+q_i)}{4} + \frac{(t-2)^2}{4m(p_i+q_i)}\right\}\Big)\right) \\
&\overset{(f)}{\le} 2\exp\left(-\Omega\Big(\min\left\{\frac{t}{2}, \frac{(t-2)^2}{2}\right\}\Big)\right) \\
&\le \exp(-\Omega(t)),
\end{aligned}
$$

where (a) uses the definition of f_i, (b) uses the fact that $\frac{a+b}{2} \le \max\{a, b\}$, inequality (c) uses the fact that $\tilde{X} + \tilde{Y} \sim \mathrm{Poi}(mp_i + mq_i)$ and the Poisson concentration bound in Theorem 21, inequality (d) uses $t \ge 3$, inequality (e) uses the fact that for $a, b > 0$, $(a+b)^2 \le a^2 + b^2$, and finally (f) uses $m \ge n$ and the fact that $x + 1/x \ge 2$ for any $x \ge 0$.

Combining the bounds on all three terms in Equation (18) proves the Lemma. ■

Finally, we prove Lemma 17.

Proof of Lemma 17 First we prove the lemma for the simpler of the two cases when $m < n$ and then later for $m \ge n$.

Proof for the regime $m < n$: Based on the value of f_i, we further divide in two cases, and for both cases we show one by one that the concentration inequality holds.

1. **Case 1:** $f_i = 1 \ge m(p_i + q_i)$.
   Since $\widehat{f_i} \ge 1$ then $\widehat{f_i} \ge f_i$, hence for any $t > 1$ we have $\Pr\left[\widehat{f_i} < \frac{1}{t} f_i\right] = 0$.

2. **Case 2:** $f_i = m(p_i + q_i) \ge 1$.
   Note for $t \ge f_i$ the inequality $\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] = 0$ trivially holds as $\widehat{f_i} \ge 1$.

For $t < f_i$

$$\begin{aligned}
\Pr\left[\widehat{f_i} < \frac{f_i}{t} f_i\right] &\overset{(a)}{=} \Pr\left[\max\{X_i + Y_i, 1\} < \frac{f_i}{t}\right] \\
&\leq \Pr\left[X_i + Y_i < \frac{f_i}{t}\right] \\
&= \Pr\left[m(p_i + q_i) - X_i + Y_i < m(p_i + q_i) - \frac{f_i}{t}\right] \\
&\overset{(b)}{=} \Pr\left[m(p_i + q_i) - X_i + Y_i < \left(1 - \frac{1}{t}\right) f_i\right] \\
&\overset{(c)}{=} \Pr\left[m(p_i + q_i) - X_i + Y_i < \frac{f_i}{2}\right] \\
&\overset{(d)}{\leq} \exp\left(-\Omega\left(\min\left\{\frac{f_i}{2}, \frac{f_i^2}{4(m(p_i + q_i)}\right\}\right)\right) \\
&\overset{(e)}{=} \exp\left(-\Omega\left(\min\left\{\frac{f_i}{2}, \frac{f_i^2}{4}\right\}\right)\right) \\
&\overset{(f)}{=} \exp(-\Omega(t))
\end{aligned}$$

where (a) uses the definition of $\widehat{f_i}$, (b) uses the fact that $f_i = m(p_i + q_i)$, inequality (c) uses $t \geq 2$, inequality (d) uses the fact that $\tilde{X} + \tilde{Y} \sim \text{Poi}(mp_i + mq_i)$ and the Poisson concentration bound in Theorem 21, inequality (e) uses $f_i = m(p_i + q_i)$, and finally (f) uses $f_i \geq t$ and $t > 1$.

Next we prove the lemma for the other case when $m \geq n$.

Proof for the regime $m \geq n$: Based on the value of f_i, we further divide in three cases, and for each of the three cases we show one by one that the concentration inequality holds.

1. **Case 1:** $f_i = 1 \geq \max\{n \cdot (p_i + q_i), \sqrt{mn} \cdot |p_i - q_i|]\}$.
   In this case $p_i + q_i \leq \frac{1}{n}$ and $|p_i - q_i| \leq \frac{1}{\sqrt{mn}}$.
   Since $\widehat{f_i} \geq 1$, then
   $$\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] \leq \Pr\left[\widehat{f_i} < f_i\right] = 0.$$

2. **Case 2:** $f_i = n \cdot (p_i + q_i) \geq \max\{1, \sqrt{mn} \cdot |p_i - q_i|]\}$.
   In this case $p_i + q_i = \frac{f_i}{n}$ and $|p_i - q_i| \leq \frac{f_1}{\sqrt{mn}}$.

Note for $t \geq f_i$ the inequality $\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] = 0$ trivially holds as $\widehat{f_i} \geq 1$. For $t < f_i$

$$\begin{aligned}
\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] &\overset{\text{(a)}}{\leq} \Pr\left[\frac{n}{m}(X_i + Y_i) \leq \frac{f_i}{t}\right] \\
&= \Pr\left[(X_i + Y_i) \leq \frac{mf_i}{nt}\right] \\
&= \Pr\left[m(p_i + q_i) - (X_i + Y_i) \geq m(p_i + q_i) - \frac{mf_i}{nt}\right] \\
&\overset{\text{(b)}}{=} \Pr\left[m(p_i + q_i) - (X_i + Y_i) \geq \frac{mf_i}{n}(1 - \frac{1}{t})\right] \\
&\overset{\text{(c)}}{\leq} \Pr\left[m(p_i + q_i) - (X_i + Y_i) \geq \frac{mf_i}{2n}\right] \\
&\overset{\text{(d)}}{\leq} 2\exp\left(- \Omega\left(\min\left\{\frac{mf_i}{2n}, \frac{m^2 f_i^2}{4n^2 \cdot m(p_i + q_i)}\right\}\right)\right) \\
&\overset{\text{(e)}}{\leq} 2\exp\left(- \Omega\left(\min\left\{\frac{mf_i}{2n}, \frac{m^2 f_i n(p_i + q_i)}{4n^2 \cdot m(p_i + q_i)}\right\}\right)\right) \\
&\leq 2\exp\left(- \Omega\left(\frac{mf_i}{4n}\right)\right) \\
&\overset{\text{(f)}}{\leq} \exp\big(- \Omega(t)\big),
\end{aligned}$$

where (a) uses the definition of $\widehat{f_i}$, (b) uses the fact that $p_i + q_i = f_i/n$ for Case 2, inequality (c) uses $t \geq 2$, inequality (d) uses the fact that $\tilde{X} + \tilde{Y} \sim \mathrm{Poi}(mp_i + mq_i)$ and the Poisson concentration bound in Theorem 21, inequality (e) uses $f_i = n(p_i + q_i)$, and finally (f) uses $m \geq n$ and $f_i > t$.

3. **Case 3:** $f_i = \sqrt{mn} \cdot |p_i - q_i| \geq \max\{1, n \cdot (p_i + q_i)\}$.
   In this case $|p_i - q_i| = f_i/\sqrt{mn}$ and $(p_i + q_i) \leq f_i/n$.

Note for $t \geq f_i$ the inequality $\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] = 0$ trivially holds as $\widehat{f_i} \geq 1$. For $t < f_i$

$$\begin{aligned}
\Pr\left[\widehat{f_i} < \frac{f_i}{t}\right] &\overset{(a)}{\leq} \Pr\left[\frac{|\tilde{X}_i - \tilde{Y}_i|}{\sqrt{m/n}} \leq \frac{f_i}{t}\right] \\
&\overset{(b)}{\leq} \Pr\left[\frac{|mp_i - mq_i| - |\tilde{X}_i - mp_i - \tilde{Y}_i + mq_i|}{\sqrt{m/n}} \leq \frac{f_i}{t}\right] \\
&= \Pr\left[\frac{|\tilde{X}_i - mp_i - \tilde{Y}_i + mq_i|}{\sqrt{m/n}} \geq \sqrt{mn}|p_i - q_i| - \frac{f_i}{t}\right] \\
&\overset{(c)}{=} \Pr\left[\frac{|\tilde{X}_i - mp_i - \tilde{Y}_i + mq_i|}{\sqrt{m/n}} \geq f_i\big(1 - \frac{1}{t}\big)\right] \\
&\overset{(d)}{\leq} \Pr\left[|\tilde{X}_i - mp_i - \tilde{Y}_i + mq_i| \geq \sqrt{\frac{m}{n}} \cdot \frac{f_i}{2}\right] \\
&\overset{(e)}{\leq} \Pr\left[\max\{|\tilde{X}_i - mp_i|, |\tilde{Y}_i - mq_i|\} \geq \sqrt{\frac{m}{n}} \cdot \frac{f_i}{4}\right] \\
&\overset{(f)}{\leq} \Pr\left[|\tilde{X}_i - mp_i| \geq \sqrt{\frac{m}{n}} \cdot \frac{f_i}{4}\right] + \Pr\left[|\tilde{Y}_i - mq_i| \geq \sqrt{\frac{m}{n}} \cdot \frac{f_i}{4}\right] \\
&\overset{(g)}{\leq} 2\exp\left(-\Omega\left(\min\left\{\sqrt{\frac{m}{n}} \cdot \frac{f_i}{4}, \frac{mf_i^2}{16n \cdot mp_i}\right\}\right)\right) \\
&\quad + 2\exp\left(-\Omega\left(\min\left\{\sqrt{\frac{m}{n}} \cdot \frac{f_i}{4}, \frac{mf_i^2}{16n \cdot mq_i}\right\}\right)\right) \\
&\overset{(h)}{\leq} \exp\big(-\Omega(t)\big),
\end{aligned}$$

where (a) uses the definition of $\widehat{f_i}$, inequality (b) follows from the triangle inequality, inequality (c) uses the fact that $\sqrt{mn}|p_i - q_i| = f_i$ for Case 3, inequality (d) uses $t \geq 2$, inequality (e) uses the fact that $|x - y| \leq 2\max\{|x|, |y|\}$, inequality (f) uses union bound, inequality (g) uses the Poisson concentration bound in Theorem 21, and finally (h) uses $m \geq n$, $f_i > n(p_i + q_i)$ and $f_i > t$.

■

Appendix C. Lower bound proofs

C.1. Proof of Theorem 7

Theorem 12 implies that for any ε_2 smaller than $\mathcal{L}(\varepsilon_1, n, m, M, \kappa, L)$, there exist random variables $U = \frac{V}{n}$ and $U' = \frac{V'}{n}$ such that

$$\begin{aligned}
&\mathbb{E}\,|U' - 1/n| \geq \varepsilon_2/n \text{ and}\\
&\mathbb{E}\,|U - 1/n| \leq \varepsilon_1/n \text{ and}\\
&\mathbb{E}\,U = \mathbb{E}\,U' = 1/n, \text{ and}\\
&\mathbb{E}\,U^i = \mathbb{E}\,U'^i, i = 2, \ldots, L, \text{ and}\\
&U, U' \in \left[\frac{(\kappa - M)}{m}, \frac{(\kappa + M)}{m}\right].
\end{aligned}$$

Next, we choose the values of parameters L, κ and M so that $\mathcal{L}(\varepsilon_1, n, m, M, \kappa, L)$ is maximized while (5) hold, which by Lemma 11 will imply $\mathrm{TV}(\mathbb{E}\,\mathrm{Poi}(mU), \mathbb{E}\,\mathrm{Poi}(mU')) \leq \frac{1}{20n}$. The choice of the parameters differs for different regimes of m.

- First we consider the regime $m < \frac{1}{4} n \log n$. Consider any such m and $\varepsilon_1 \leq 1/8$. Choose $\kappa = M = \log n$ and $L = 4e^2 \log n$. One can check that the desired bound on TV distance in (5) is satisfied for these choices of the parameters. Further, we have $A = \frac{2n \log n}{m} - 1 - \varepsilon_1 \geq \frac{n \log n}{m}$, where we used $\frac{n \log n}{m} > 4 > 1 + \varepsilon_1$ in the above parameter range; and $B = 1 - \varepsilon_1 \geq 1/2$. Finally, $\varepsilon_1 < 1/8$ we have that $\varepsilon_1 \leq \min\left\{\frac{B}{4}, \frac{A}{4}\right\}$. Then, invoking Theorem 10 we get that for any
$$\varepsilon_2 \leq \mathcal{L}(\varepsilon_1, n, m, M = \log n, \kappa = \log n, L = 4e^2 \log n)$$
and $\varepsilon_2^2 \geq 1000\frac{M}{m} = 1000\frac{\log n}{m}$, one cannot distinguish between $25\varepsilon_1$-close and $\varepsilon_2/2$-far using $m/2$ samples.

  Equivalently, by rescaling the parameters, for any $m < c_0 n \log n$,
$$\varepsilon_2 \leq \frac{1}{2}\mathcal{L}\Big(\frac{\varepsilon_1}{25}, n, 2m, M = \log n, \kappa = \log n, L = 4e^2 \log n\Big), \tag{19}$$
and $\frac{\varepsilon_2^2}{2^2} \geq 1000\frac{\log n}{m}$, we can not distinguish between ε_1-close and ε_2-far using m samples.

  Next, we show that the above statement holds even without the constrain $\frac{\varepsilon_2^2}{2^2} \geq 1000\frac{\log n}{m}$. We do so by showing that even when $\frac{\varepsilon_2^2}{2^2} < 1000\frac{\log n}{m}$ or equivalently $m < 4000\frac{\log n}{\varepsilon_2^2}$, we can not distinguish between ε_1-close and ε_2-far using m samples, even for $\varepsilon_1 = 0$. From the known uniformity testing lower bound for non tolerant case, we know that there is an absolute constant $c_3 > 0$ such that $m \geq c_3\frac{\sqrt{n}}{\varepsilon_2^2}$ samples are needed to distinguish correctly between $p = \mathrm{Unif}_n$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$ with probability $\geq 4/5$. Since for n larger than an absolute constant c_4, we have $c_3\frac{\sqrt{n}}{\varepsilon_2^2} > 4000\frac{\log n}{\varepsilon_2^2} > m$, hence m samples are insufficient and the claim follows.

  From Theorem 12 we get:
$$\frac{1}{2}\mathcal{L}\Big(\frac{\varepsilon_1}{25}, n, 2m, M = \log n, \kappa = \log n, L = 4e^2 \log n\Big) \geq \max\left\{c_1\sqrt{\frac{\varepsilon_1 n}{m \log n}}, c_2\sqrt{\frac{\varepsilon_1 \sqrt{n}}{\sqrt{m \log n}}}\right\},$$

where c_1 and c_2 are some absolute positive constants.

From (19) and the above lower bound on $\mathcal{L}$ it follows that for any $n > c_4$, for any $m < (n \log n)/4$ and $\varepsilon_1 < 1/8$, such that

$$m < \max\left\{ c_1^2 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right), c_2^3 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 \right\},$$

then using m samples from p one can not distinguish correctly with probability $\geq 4/5$ between $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$.

Observe given c_1 and c_2, there exist an universal constant c_5 such that for any $\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) \leq c_5 \log n$, we have

$$\max\left\{ c_1^2 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right), c_2^3 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 \right\} \leq \frac{n \log n}{4}.$$

Therefore, for any $\varepsilon_1 \leq \frac{1}{8}$, $n > c_4$ and $\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) \leq c_5 \log n$, then using $\max\left\{ c_1^2 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right), c_2^3 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 \right\} = \Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 \right)$ samples from p one can not distinguish correctly with probability $\geq 4/5$ between $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$.

- Next, we choose the parameters for the regime $m > 4n \log n$. Note that since the theorem statement makes no claim for the setting where $m \geq \frac{n}{\varepsilon_2^2 \log n}$, we can restrict our attention to the case where $m \leq \frac{n}{\varepsilon_2^2 \log n}$. Furthermore, because $\frac{n}{\varepsilon_2^2 \log n} \ll \frac{n \log n}{64 \varepsilon_1^2}$ (using the fact that $\varepsilon_1 \leq \varepsilon_2$), we only need to consider $m \leq \frac{n \log n}{64 \varepsilon_1^2}$. This condition on m implies that $\varepsilon_1 \leq \frac{1}{8}\sqrt{\frac{n \log n}{m}}$. Consider any such m. Choose $\kappa = \frac{m}{n}$ and $M = \sqrt{\frac{m \log n}{n}}$ and $L = 4e^2 \log n$. Observe that for this choice $M < \frac{m}{n} = \kappa$, hence $\kappa - M > 0$. Then $A = B = \frac{nM}{m} - \varepsilon_1 = \sqrt{\frac{n \log n}{m}} - \varepsilon_1$. Since $\varepsilon_1 \leq \frac{1}{8}\sqrt{\frac{n \log n}{m}}$, observe that $A, B \geq \frac{1}{2}\sqrt{\frac{n \log n}{m}}$ and $\varepsilon_1 \leq \min\left\{\frac{B}{4}, \frac{A}{4}\right\}$. Then, invoking Theorem 10 and rescaling the parameters as for the previous case, we get

  $$\varepsilon_2 \leq \frac{1}{2}\mathcal{L}\left(\frac{\varepsilon_1}{25}, n, 2m, M = \sqrt{\frac{m \log n}{n}}, \kappa = \frac{m}{n}, L = 4e^2 \log n \right),$$

  and $\frac{\varepsilon_2^2}{2^2} \geq 1000 \frac{M}{m} = 1000\sqrt{\frac{\log n}{mn}}$, we can not distinguish between ε_1-close vs ε_2-far using m samples. From Theorem (12) we get:

  $$\mathcal{L}\left(\varepsilon_1, n, m, M = \sqrt{\frac{m \log n}{n}}, \kappa = \frac{m}{n}, L = 4e^2 \log n \right) \geq c_6 \sqrt{\varepsilon_1 \cdot \frac{\sqrt{n \log n}}{\sqrt{m} \log n}},$$

  where c_6 is some absolute constant.

  Following the similar steps as before it can be shown that for any $\varepsilon_1 \leq \frac{1}{8}$, $n > c_8$ and $\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) \geq c_7 \log n$, then using $c_6^4 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2$ samples from p one can not distinguish correctly with probability $\geq 4/5$ between $\|p - \mathrm{Unif}_n\|_1 \leq \varepsilon_1$ and $\|p - \mathrm{Unif}_n\|_1 \geq \varepsilon_2$. Using $\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) \geq c_7 \log n$, we get $c_6^4 \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 = \Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right)$ bound on the sample complexity.

We have so far shown the target lower bound on the sample complexity for both regimes $\frac{\varepsilon_1}{\varepsilon_2^2} < c_5 \log n$ and $\frac{\varepsilon_1}{\varepsilon_2^2} > c_7 \log n$ for some absolute positive constants c_5 and c_7. To conclude for the intermediate cases, observe that the sample complexity is an increasing function of ε_1 (more tolerance makes the problem harder) and a decreasing function of ε_2. Thus, by monotonicity, the lower bound for $\frac{\varepsilon_1}{\varepsilon_2^2} = c_5 \log n$ still applies to $c_5 \log n \leq \frac{\varepsilon_1}{\varepsilon_2^2} \leq c_7 \log n$, by relaxing the problem to $\varepsilon_1' = \frac{c_5 \varepsilon_1}{c_7}$. This only affects the resulting lower bound by a constant factor, and allows us to conclude with the desired

$$\Omega\left(\frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \frac{n}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)\right)$$

sample complexity lower bound for the full range of parameters.

C.2. Proof of Theorem 12

We break the proof of Theorem 12 into two parts. First, we convert the primal form of the problem into a more convenient representation via a few helpful transformations, and then take the dual (Section C.2.1). We then lower bound the value of the dual using tools from approximation theory (Section C.2.2).

C.2.1. Transforming the primal

First, to simplify the optimization problem (6), notice that moment matching of all degree-L or less moments is unaffected by translation. So if one introduces the random variables $X = V - 1$ and $X' = V' - 1$, we see that these are supposed to be mean zero random variables over $[\frac{n(\kappa - M)}{m} - 1, \frac{n(\kappa+M)}{m} - 1]$ with matching L-th and below moments, and that distance to uniformity corresponds to $\mathbb{E}\,|X|$ and $\mathbb{E}\,|X'|$.

$$\begin{aligned} \max \mathbb{E}\,|X'| \text{ s.t. } & \mathbb{E}\,|X| \leq \varepsilon_1 \text{ and} \\ & \mathbb{E}\,X = \mathbb{E}\,X' = 0 \\ & \mathbb{E}\,X^i = \mathbb{E}\,X'^i, i = 2, \ldots, L, \text{ and} \\ & X, X' \in \left[\frac{n(\kappa - M)}{m} - 1, \frac{n(\kappa + M)}{m} - 1\right]. \end{aligned} \tag{20}$$

It will be useful to remove the constraint that $\mathbb{E}[X] = \mathbb{E}[X'] = 0$. To do so, we propose the following optimization problem without this constraint.

$$\begin{aligned} \max \mathbb{E}\,|Y'| \text{ s.t. } & \mathbb{E}\,|Y| \leq \frac{\varepsilon_1}{2} \text{ and} \\ & \mathbb{E}\,Y^i = \mathbb{E}\,Y'^i, i = 1, \ldots, L, \text{ and} \\ & Y, Y' \in [-B, A], \end{aligned} \tag{21}$$

where $A = \frac{n(\kappa+M)}{m} - 1 - \varepsilon_1$ and $B = -\left(\frac{n(\kappa-M)}{m} - 1 + \varepsilon_1\right)$, first defined in Equation (7). We show the following claim.

Lemma 22 *The value of the solution of* (20) *is at least half the value of the solution of* (21)*.*

Proof Let Y and Y' be the random variables that achieve the maximum in (21). To prove the claim, first we show that random variables $X = Y - \mathbb{E}\,Y$ and $X' = Y' - \mathbb{E}\,Y$ satisfy the constrains in (20).

First note that $\mathbb{E}\,|X| = \mathbb{E}\,|Y - \mathbb{E}\,Y| \le 2\,\mathbb{E}\,|Y| \le \varepsilon_1$. Next, using $\mathbb{E}\,Y = \mathbb{E}\,Y'$, we get $\mathbb{E}\,X = \mathbb{E}[Y - \mathbb{E}\,Y] = 0$ and $\mathbb{E}\,X' = \mathbb{E}[Y' - \mathbb{E}\,Y] = 0$. Since the moment matching of all degree-L or less moments is unaffected by translation, all L moments of X and X' will match. Finally, $|\,\mathbb{E}[Y]| \le \mathbb{E}[|Y|] \le \frac{\varepsilon_1}{2}$, and $Y, Y' \in [\frac{n(\kappa-M)}{m} - 1 + \varepsilon_1, \frac{n(\kappa+M)}{m} - 1 - \varepsilon_1]$ ensures $X, X' \in [\frac{n(\kappa-M)}{m} - 1, \frac{n(\kappa+M)}{m} - 1]$.

Now to complete the proof we show that the solution of the optimization problem (20) is at least $\ge \frac{\mathbb{E}\,|Y'|}{2}$.

If $\mathbb{E}\,|Y'| \le \varepsilon_1$, then this is obviously true as the maximum in (20) is at least ε_1. If $\mathbb{E}\,|Y'| \ge \varepsilon_1$ then $\mathbb{E}\,|X'| = \mathbb{E}\,|Y' - \mathbb{E}\,Y| \ge \mathbb{E}\,|Y'| - \mathbb{E}\,|Y| \ge \mathbb{E}\,|Y'| - \frac{\varepsilon_1}{2} > \frac{\mathbb{E}\,|Y'|}{2}$, where we used $\mathbb{E}\,|Y'| \ge \varepsilon_1$ in the last step. ■

Mechanically taking the dual of (21), we obtain that the dual is:

$$\begin{aligned} \min \frac{\varepsilon_1}{2}\alpha + z_1 + z_2 \text{ s.t. } & z_1 + \sum_{i=1}^{L} c_i x^i \ge |x| \text{ for all } x \in [-B, A], \\ & \alpha\,|x| \ge \sum_{i=1}^{L} c_i x^i - z_2 \text{ for all } x \in [-B, A], \text{ and} \\ & \alpha \ge 0\,. \end{aligned} \tag{22}$$

By weak duality, we know that the value of the optimal solution to (21) is upper bounded by the value of the optimal solution to (22). However, since we seek to prove a lower bound on the value of the optimal solution to (21), this is insufficient for our purposes. However, we show that in this case, strong duality still holds:

Lemma 23 *The value of the optimal solution to* (21) *is equal to the value of the optimal solution to* (22)*.*

This follows from the classical theory of convex duality Rockafellar (1974), however, for completeness we also give a self-contained proof of this fact in Section C.5.

We now make a couple of final simplifications before we lower bound the value of (22). Let $\mathcal{P}_L$ denote the collection of all degree-L polynomials. We reparametrize the above dual by letting $\varepsilon_2 \coloneqq \alpha\varepsilon_1$, $p(x) \coloneqq \frac{\varepsilon_1}{\varepsilon_2}\sum_{i=1}^{L} c_i x^i - z_2$ and $z \coloneqq z_1 + z_2$.

$$\begin{aligned} \min \frac{\varepsilon_2}{2} + z \text{ s.t. } & \text{for some } p(x) \in \mathcal{P}_L, \\ & p(x) \ge \frac{\varepsilon_1}{\varepsilon_2}\,|x| - \frac{\varepsilon_1}{\varepsilon_2} z \text{ for all } x \in [-B, A], \text{ and} \\ & |x| \ge p(x) \text{ for all } x \in [-B, A], \text{ and} \\ & \varepsilon_2 \ge 0. \end{aligned} \tag{23}$$

Imposing the additional constraint that $z = \varepsilon_2$, we get the following program:

$$\begin{aligned} \min \varepsilon_2 \text{ s.t. } & \text{for some } p(x) \in \mathcal{P}_L, \\ & p(x) \geq \frac{\varepsilon_1}{\varepsilon_2} |x| - \varepsilon_1 \text{ for all } x \in [-B, A], \text{ and} \\ & |x| \geq p(x) \text{ for all } x \in [-B, A], \text{ and} \\ & \varepsilon_2 \geq 0. \end{aligned} \tag{24}$$

Suppose $\varepsilon_2 = a$ and $z = b$ achieve the optimal solution of (23), which is therefore $a/2+b$. It is easy to verify that for $z < 0$ the optimization problem (23) is infeasible, hence $b \geq 0$. Further, observe that $z = \varepsilon_2 = \max\{a, b\}$ is also a feasible solution of (23), and $z + \varepsilon_2/2 = \max\{3a/2, 3b/2\}$, which is at most 3 times the optimal solution of (23). Since (24) scales it down by $2/3$, it follows that the solution of the new dual (24) is at most twice the solution of the previous dual (23).

Lemma 24 *The value of the solution of* (6) *is at least* $1/4$ *times the value of the solution of* (24).

Proof From the preceding discussion, the value of the solution of (6) is equal to the value of the solution of (20). From Lemma 22, this value is at least is at least $1/2$ times the value of the solution of (21). From Lemma 23 the values of the solutions of (21) and (22) are equal. Furthermore, the value of the solution of (22) is equal to the value of the solution of (23). Finally, the value of the solution of (23) is at least $1/2$ times the value of the solution of (24). Hence, the value of the solution of (6) is at least $1/4$ times the value of the solution of (24). ■

C.2.2. LOWER BOUNDING THE DUAL

We now establish a lower bound on the solution of the dual (Equation (24)). We assume that parameters A, B are such that $\varepsilon_1 \ll A, B$. First note that when $\varepsilon_2 \leq \frac{\varepsilon_1}{2}$, then the two conditions are contradictory and can not be met simultaneously. Therefore, we get the lower bound:

$$\varepsilon_2 \geq \frac{\varepsilon_1}{2}. \tag{25}$$

This implies that $\forall\, x \in [-B, A]$

$$p(x) \geq 2|x| - \varepsilon_1.$$

Combining this with constraints in the dual imply that for all $x \in [-B, A]$

$$|p(x)| \leq 2|x| + \varepsilon_1.$$

From the above equation, for $x = 0$, we get $|p_0| \leq \varepsilon_1$. Then $\forall\, x \in [-B, A]$,

$$\begin{aligned} |p(x) - p_0| &\leq |p(x)| + |p_0| \\ &\leq 2|x| + \varepsilon_1 + \varepsilon_1 \\ &= 2|x| + 2\varepsilon_1. \end{aligned}$$

Let $p(x) \in \mathcal{P}_L$ be any polynomial satisfying the constrains of the dual. Let $p_0 = p(0)$, p_1 be the coefficients of x in $p(x)$, and $\tilde{p}(x) = p(x) - p_0 - p_1 x$.

Using the above equation,

$$|\tilde{p}(x)| \leq |p(x) - p_0| + |p_1 x| \leq (2 + |p_1|)|x| + 2\varepsilon_1. \tag{26}$$

The constraints in the dual also imply $p_0 \leq 0$ and $p(3\varepsilon_2) \geq 2\varepsilon_1$. Combining these,

$$p(2\varepsilon_2) - p_0 \geq 2\varepsilon_1.$$

Using $\tilde{p}(x) = p(x) - p_1 x - p_0$ for $x = 3\varepsilon_2$ in the above equation

$$\tilde{p}(3\varepsilon_2) \geq 2\varepsilon_1 - p_1 3\varepsilon_2.$$

Similarly, one can get

$$\tilde{p}(-3\varepsilon_2) \geq 2\varepsilon_1 + p_1 3\varepsilon_2.$$

Combining the two equations we get

$$\max\{\tilde{p}(3\varepsilon_2), \tilde{p}(-3\varepsilon_2)\} \geq 2\varepsilon_1 + |p_1| 3\varepsilon_2. \tag{27}$$

Our lower bound on the dual is the consequence of the following key lemma, which we prove in Section C.3.

Lemma 25 *Let g be a degree-L polynomial such that $g(0) = g'(0) = 0$. For some $a < 0 < b$, $\gamma \geq 1$ and $0 \leq \delta \leq \min\left\{-\frac{a}{4}, \frac{b}{4}\right\}$ suppose $|g(x)| \leq \gamma|x| + \delta$ for all $x \in [a, b]$ then*

$$\min\{|x| : g(x) \geq (\gamma - 1)|x| + \delta\} \geq \begin{cases} \sqrt{\delta \cdot \frac{b-a}{32L^2}} & \text{if } \delta < \frac{(b-a)}{32L^2} \\ \sqrt{\delta \cdot \frac{\sqrt{|ab|}}{16L}} & \text{if } \delta < \frac{\sqrt{|ab|}}{16L}. \end{cases}$$

The next theorem that establishes the lower bound on the dual follows by combining Equation (27), Equation (26) and Lemma 25.

Theorem 26 (Lower bound on the solution to the dual) *For any $A, B > 0$ and $0 < \varepsilon_1 \leq \min\left\{\frac{B}{4}, \frac{A}{4}\right\}$, the value of optimal solution of (24) is lower bounded by*

$$\varepsilon_2 \geq \max\left\{\frac{1}{3}\sqrt{\varepsilon_1 \cdot \frac{A+B}{32L^2}}, \frac{1}{3}\sqrt{\varepsilon_1 \cdot \frac{\sqrt{AB}}{16L}}, \frac{\varepsilon_1}{2}\right\}$$

Proof Applying Lemma 25 for $g = \tilde{p}/2$, $\gamma = 1 + |p_1|/2$, $\delta = \varepsilon_1$, $b = A$, and $a = -B$ gives

$$3\varepsilon_2 \geq \begin{cases} \sqrt{\varepsilon_1 \cdot \frac{A+B}{32L^2}} & \text{if } \varepsilon_1 < \frac{(A+B)}{32L^2} \\ \sqrt{\varepsilon_1 \cdot \frac{\sqrt{AB}}{16L}} & \text{if } \varepsilon_1 < \frac{\sqrt{AB}}{16L}. \end{cases}$$

Combining the above bound with the upper bound $\varepsilon_2 \geq \frac{\varepsilon_1}{2}$ in (25), and using the observations that if $\varepsilon_1 \geq \frac{(A+B)}{32L^2}$ then $\sqrt{\varepsilon_1 \cdot \frac{A+B}{32L^2}} \leq \varepsilon_1$, and if $\varepsilon_1 \geq \frac{\sqrt{AB}}{16L}$ then $\sqrt{\varepsilon_1 \cdot \frac{\sqrt{AB}}{16L}} \leq \varepsilon_1$, completes the proof. ■

Combining Lemma 24 and Theorem 26 proves Theorem 12.

C.3. Proof of Lemma 25

First, we recall some useful results from approximation theory, which we then leverage to derive a few auxiliary lemmas. Finally, using these lemmas we establish Lemma 25.

We will use the two following results, both of which bound the absolute value of derivatives of bounded polynomials. The first is the Markov Brothers' inequality, which gives a bound on all derivatives of bounded polynomials.

Theorem 27 (Markov Brothers' inequality Markoff (1892)) *For any real polynomial g of degree-L, $k \geq 0$, and real numbers a and b s.t. $-\infty < a < b < \infty$, the k^{th} derivative of g satisfies*

$$\max_{x\in[a,b]} |g^{(k)}(x)| \leq 2^k \cdot \frac{\max_{x\in[a,b]} |g(x)|}{(b-a)^k} \cdot \prod_{i=0}^{k-1} \frac{(L^2-i^2)}{2i+1}.$$

We next need Bernstein's inequality, which provides a bound on the first derivative, which for some values of x is stronger than the Markov Brothers' inequality.

Theorem 28 (Bernstein's inequality Bernstein (1912)) *For any real polynomial g of degree-L and real numbers $-\infty < a < x < b < \infty$, we have*

$$|g'(x)| \leq \frac{L \cdot \max_{x\in[a,b]} |g(x)|}{\sqrt{(x-b)(a-x)}}.$$

Using Bernstein's inequality above we can obtain the following bounds on the first and the second derivatives, which for some range of parameters improve upon Markov Brothers' inequality. Note that in the following lemma we have assumed $a < 0 < b$.

Lemma 29 *For any real polynomial g of degree-L, and real numbers a and b s.t. $-\infty < a < 0 < b < \infty$, the following hold*

$$\max_{x\in[a/2,b/2]} |g'(x)| \leq \frac{L \cdot \max_{x\in[a,b]} |g(x)|}{\sqrt{|ba|/2}},$$

and

$$\max_{x\in[a/4,b/4]} |g''(x)| \leq \frac{4L(L-1) \cdot \max_{x\in[a,b]} |g(x)|}{|ab|}.$$

Proof From Bernstein's inequality we get

$$\max_{x\in[a/2,b/2]} |g'(x)| \leq \frac{L \cdot \max_{x\in[a,b]} |g(x)|}{\sqrt{|ba|/2}},$$

where we used $(x-b)(a-x) \geq |ab|/2$ if $a < 0 < b$ and $x \in [a/2, b/2]$.
Replacing $g \to g'$, $a \to a/2$ and $b \to b/2$ in the above equation, we get

$$\max_{x\in[a/4,b/4]} |g''(x)| \leq \frac{(L-1) \cdot \max_{x\in[a/2,b/2]} |g'(x)|}{\sqrt{|ba|/8}}.$$

Combining the two equations proves the lemma. ■

Using the bounds on the derivative of the bounded functions in Theorem 27 and Lemma 29, and a Taylor expansion, we derive the following lemma.

Lemma 30 *Let g be a degree-L polynomial such that $g(0) = g'(0) = 0$. Suppose that, for some $a < 0 < b$, we have $|g(x)| \leq |x|$ for all $x \in [a, b]$. Then the following bounds hold:*

$$|g(x)| \leq \begin{cases} \frac{8L^2}{(b-a)}x^2 & \textit{if } x \in [a/4, b/4] \textit{ and } |x| \leq \frac{3(b-a)}{L^2} \\ \frac{4L}{\sqrt{|ab|}}x^2 & \textit{if } x \in [a/4, b/4] \textit{ and } |x| \leq \frac{\sqrt{|ab|}}{4L}. \end{cases}$$

Proof By Taylor's theorem, we know that

$$g(x) = g(0) + g'(0)x + \frac{g''(c)}{2}x^2,$$

where c is some number between 0 and x.
Then using $g(0) = g'(0) = 0$, we obtain

$$|g(x)| \leq x^2 \cdot \max_{z \leq |x|} \left|\frac{g''(z)}{2}\right|. \tag{28}$$

Next, let $h(x) = g(x)/x$. Note that $h(x)$ is a degree-$(L-1)$ polynomial and $\max_{[a,b]} |h(x)| \leq 1$. From Theorem 27 and Lemma 29 it follows

$$\max_{x \in [a/2, b/2]} |h'(x)| \leq \min\left\{\frac{2L^2}{b-a}, \frac{\sqrt{2} \cdot L}{\sqrt{|ab|}}\right\},$$

and

$$\max_{x \in [a/4, b/4]} |h''(x)| \leq \min\left\{\frac{4L^4}{3(b-a)^2}, \frac{4L^2}{|ab|}\right\}.$$

The following relation between the second derivative of g and the derivatives of h can be obtained by differentiation $x \cdot h(x)$ twice using the chain rule:

$$g''(x) = 2h'(x) + xh''(x).$$

Using the bounds on the derivatives of h, for any $x \in [a/4, b/4]$,

$$|g''(x)| \leq \min\left\{\frac{4L^2}{(b-a)} + |x|\frac{4L^4}{3(b-a)^2}, \frac{2\sqrt{2} \cdot L}{\sqrt{|ab|}} + |x|\frac{4L^2}{|ab|}\right\}.$$

Note that for any $x \in [a/4, b/4]$ s.t. $|x| \leq \frac{3(b-a)}{L^2}$,

$$|g''(x)| \leq \frac{8L^2}{(b-a)},$$

and similarly for any $x \in [a/4, b/4]$ s.t. $|x| \leq \frac{\sqrt{|ab|}}{4L}$,

$$|g''(x)| \leq \frac{4L}{\sqrt{|ab|}}.$$

Combining the above bounds with Equation (28) proves the lemma. ■

Using this, we derive the following lemma.

Lemma 31 *Let g be a degree-L polynomial such that $g(0) = g'(0) = 0$. Supoose that for some $a < 0 < b$, $\gamma \geq 1$ and $0 < \delta \leq \min\left\{-\frac{a}{4}, \frac{b}{4}\right\}$ such that*

$$\delta < \max\left\{\frac{(b-a)}{16L^2}, \frac{\sqrt{|ab|}}{8L}\right\},$$

we have $|g(x)| \leq \gamma|x| + \delta$ for all $x \in [a,b]$. Then $g(x) \leq 2\gamma|x|$ for all $x \in [a,b]$.

Proof Note that $|g(x)| \leq \gamma|x| + \delta$ implies that $|g(x)| \leq 2\gamma|x|$ for $|x| \geq \frac{\delta}{\gamma}$. We proceed by contradiction. For contradiction, assume

$$\max_{|x|\leq\frac{\delta}{\gamma}} \frac{|g(x)|}{|x|} = r \geq 1.$$

Consider the polynomial $\tilde{g}(x) \coloneqq g(x)/r$. Then from the definition of r, for any x such that $|x| \leq \frac{\delta}{\gamma}$ it follows that

$$|\tilde{g}(x)| = |g(x)|/r \leq |x|, \tag{29}$$

and

$$|\tilde{g}(x)| = |x| \tag{30}$$

for at least one x such that $|x| \leq \frac{\delta}{\gamma}$. To show the contradiction, in the reminder of the proof, we show that this is impossible. For $x \in [a,b] \setminus [-\frac{\delta}{\gamma}, \frac{\delta}{\gamma}]$,

$$|\tilde{g}(x)| = \frac{|g(x)|}{r} \leq |g(x)| \leq \gamma|x| + \delta \leq 2\gamma|x|,$$

where the last inequality uses the fact that for $|x| \geq \frac{\delta}{\gamma}$ we get $\delta \leq |x|$. Combining with Equation (29), we get $|\tilde{g}(x)| \leq 2\gamma|x|$ for all $x \in [a,b]$.
Applying Lemma 30 on $\frac{\tilde{g}(x)}{2\gamma}$, we get

$$|\tilde{g}(x)| \leq \begin{cases} 2\gamma\frac{8L^2}{(b-a)}x^2 & \text{if } x \in [a/4, b/4] \text{ and } |x| \leq \frac{3(b-a)}{L^2} \\ 2\gamma\frac{4L}{\sqrt{|ab|}}x^2 & \text{if } x \in [a/4, b/4] \text{ and } |x| \leq \frac{\sqrt{|ab|}}{4L}. \end{cases}$$

Recall $0 \leq \delta \leq \min\left\{-\frac{a}{4}, \frac{b}{4}\right\}$ and $\gamma \geq 1$. First we show a contradiction when $\delta < \frac{b-a}{16L^2}$. In this case, $\frac{\delta}{\gamma} \leq \delta \leq \frac{b-a}{16L^2} < \frac{3(b-a)}{L^2}$, hence for any x such that $|x| \leq \frac{\delta}{\gamma}$ the above equation implies that

$$|\tilde{g}(x)| \leq 2\gamma\frac{8L^2}{(b-a)}|x|^2 \leq \gamma\frac{16L^2}{(b-a)} \cdot |x| \cdot \max|x| \leq \gamma\frac{16L^2}{(b-a)} \cdot \frac{\delta}{\gamma} = \frac{16L^2}{(b-a)} \cdot |x| \cdot \delta < |x|,$$

where the last inequality uses $\delta < \frac{b-a}{16L^2}$. This contradicts Equation (30).

Next we show a contradiction when $\delta < \frac{\sqrt{|ab|}}{8L}$ by using similar steps. In this case, $\frac{\delta}{\gamma} \leq \delta \leq \frac{\sqrt{|ab|}}{8L} < \frac{\sqrt{|ab|}}{4L}$, hence for any x such that $|x| \leq \frac{\delta}{\gamma}$ we get

$$|\tilde{g}(x)| \leq 2\gamma\frac{4L}{\sqrt{|ab|}}|x|^2 \leq \gamma\frac{8L}{\sqrt{|ab|}} \cdot |x| \cdot \max|x| \leq \gamma\frac{8L}{\sqrt{|ab|}} \cdot \frac{\delta}{\gamma} = \gamma\frac{4L}{\sqrt{|ab|}} \cdot |x| \cdot \delta < |x|,$$

where the last inequality uses $\delta < \frac{\sqrt{|ab|}}{8L}$. This contradicts Equation (30). ■

Finally, we use Lemmas 30 and 31 to derive Lemma 25.

Proof of Lemma 25. First consider the case $\delta < \frac{(b-a)}{16L^2}$. Then Lemma 31 implies $g(x) \leq 2\gamma|x|$. Applying Lemma 30 gives

$$|g(x)| \leq 2\gamma \frac{8L^2}{(b-a)} x^2.$$

To prove the first bound in the lemma we show that if $|x| < \sqrt{\delta \cdot \frac{b-a}{32L^2}}$ then $g(x) < (\gamma - 1)|x| + \delta$. First we show it when $\gamma \leq 2$. In this case, for all x such that $|x| < \sqrt{\delta \cdot \frac{b-a}{32L^2}}$,

$$|g(x)| \leq 2\gamma \frac{8L^2}{(b-a)} x^2 < \frac{\gamma}{2}\delta < \delta.$$

Then, we turn to the case $\gamma \geq 2$. In this case, for all x such that $|x| < \sqrt{\delta \cdot \frac{b-a}{32L^2}}$,

$$|g(x)| \leq 2\gamma \frac{8L^2}{b-a} x^2 < \frac{\gamma}{2}|x| < (\gamma - 1)|x|,$$

where step the second inequality uses $|x| < \frac{b-a}{32L^2}$, which follows since $|x| < \sqrt{\delta \frac{b-a}{32L^2}}$ and $\delta < \frac{b-a}{32L^2}$, and the last inequality uses $\frac{\gamma}{2} < (\gamma - 1)$, since $\gamma \geq 2$. This completes the proof of the first upper bound in the lemma. The second upper bound in the lemma can be shown using similar steps. □

C.4. Proof of Theorem 10

Let $\mathcal{P}^{(1)}$ be the joint distribution of n independent copies of U and $\mathcal{P}^{(2)}$ be the joint distribution of n independent copies of U', respectively. For $j = 1, 2$, let $(U_1^{(j)}, \ldots, U_n^{(j)}) \sim \mathcal{P}^{(j)}$. Define random vectors $D^{(j)}$ as the ℓ_1-normalizations of these vectors:

$$D^{(j)} := \left(\frac{U_1^{(j)}}{\sum_{i=1}^n U_i^{(j)}}, \ldots, \frac{U_n^{(j)}}{\sum_{i=1}^n U_i^{(j)}} \right).$$

Since $U_i^{(j)} \geq 0$, the vectors $D^{(j)}$ are distributions.

Let $N^{(j)} \sim \text{Poi}(m \sum_i U_i^{(j)})$. Let $(C_1^{(j)}, \ldots, C_n^{(j)})$ be the collection of random variables, whose joint distribution conditioned on $(U_1^{(j)}, \ldots, U_n^{(j)})$ and $N^{(j)}$ is the following multinomial distribution,

$$(C_1^{(j)}, \ldots, C_n^{(j)}) \Big| \left(\left(U_1^{(j)}, \ldots, U_n^{(j)} \right), N^{(j)} \right) \sim \text{Mult}\left(\left(\frac{U_1^{(j)}}{\sum U_i^{(j)}}, \ldots, \frac{U_n^{(j)}}{\sum U_i^{(j)}} \right), N^{(j)} \right) \equiv \text{Mult}\left(D^{(j)}, N^{(j)} \right).$$

It follows that we can use $(C_1^{(j)}, \ldots, C_n^{(j)})$ to generate up to $N^{(j)}$ samples from $D^{(j)}$. Observe that for $i \in [n]$, conditioned on the $U_i^{(j)}$'s, $C_i^{(j)} \sim \text{Poi}(mU_i^{(j)})$ are independent Poisson random variables.

Define the events

$$E^{(1)} = \left\{\left|\sum_{i=1}^{n} U_i^{(1)} - 1\right| \le \frac{1}{10}\right\} \bigcap \left\{\sum_{i=1}^{n}\left|U_i^{(1)} - \frac{1}{n}\right| \le 10\varepsilon_1\right\},$$

and

$$E^{(2)} = \left\{\left|\sum_{i=1}^{n} U_i^{(2)} - 1\right| \le \frac{\varepsilon_2}{10}\right\} \bigcap \left\{\sum_{i=1}^{n}\left|U_i^{(2)} - \frac{1}{n}\right| \ge \frac{9\varepsilon_2}{10}\right\}.$$

We bound the probability of the complement events $\overline{E^{(1)}}$ and $\overline{E^{(2)}}$. The following general lemma will be useful, which we prove using Chebyshev's inequality.

Lemma 32 *Let $X_1, \ldots, X_n$ be n i.i.d. random variables over $[a, b]$ for some $0 \le a < b$ and $\mathbb{E}[X_i] = \mu$. Then*

$$\Pr\left[\left|\sum_i X_i - n\mu\right| \ge \sqrt{10n\mu(b-a)}\right] \le \frac{1}{10}$$

Proof The following bound on the variance of a random variable will be useful. This bound has been proved in many previous works including Bhatia and Davis (2000). We provide the proof for completeness.

Theorem 33 *Let X be any random variable over $[a, b]$, then $\mathrm{Var}(X) \le (b - \mathbb{E}[X])(\mathbb{E}[X] - a)$.*

Proof Let $Y = \frac{X-a}{b-a}$. Note $Y \in [0, 1]$. Then

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \le \mathbb{E}[Y] - (\mathbb{E}[Y])^2 = \mathbb{E}[Y](1 - \mathbb{E}[Y]),$$

where we used the fact $Y^2 \le Y$, since $Y \in [0, 1]$. Then using the relations $\mathrm{Var}(Y) = \frac{\mathrm{Var}(X)}{(b-a)^2}$ and $\mathbb{E}[Y] = \frac{\mathbb{E}[X]-a}{b-a}$ completes the proof. ■

From Theorem 33

$$\mathrm{Var}(X_i) \le (b - \mu)(\mu - a) \le (b - a)\mu,$$

where we used $0 \le a \le \mu \le b$. Then $\mathrm{Var}(\sum X_i) \le n\mu(b - a)$. From Chebyshev's inequality

$$\Pr\left[\left|\sum_i X_i - \sum_i \mathbb{E}[X_i]\right| \ge \sqrt{10n\mu(b-a)}\right] \le \frac{1}{10}.$$

■

First we bound the probability of $\overline{E^{(1)}}$. Using the union bound

$$\Pr[\overline{E^{(1)}}] \le \Pr\left[\left|\sum_{i=1}^{n} U_i^{(1)} - \frac{1}{n}\right| \ge \frac{1}{10}\right] + \Pr\left[\sum_{i=1}^{n}\left|U_i^{(1)} - \frac{1}{n}\right| \ge 10\varepsilon_1\right].$$

We next upper bound both these terms, starting with the former. Note that $U_i^{(1)} \in [a,b]$ and $\mathbb{E}\left[U_i^{(1)}\right] = \frac{1}{n}$. Applying Lemma 32, we obtain

$$\Pr\left[\left|\sum_i U_i^{(1)} - 1\right| \geq \sqrt{10(b-a)}\right] \leq \frac{1}{10},$$

which, since $10(b-a) \leq \frac{\varepsilon_2^2}{100} \leq \frac{1}{100}$, upper bounds the first term in the expression.
We now bound the second term using linearity of expectations and Markov's inequality,

$$\Pr\left[\sum_{i=1}^n \left|U_i^{(1)} - \frac{1}{n}\right| \geq 10\varepsilon_1\right] \leq \frac{\sum_i \mathbb{E}\left[\left|U_i^{(1)} - \frac{1}{n}\right|\right]}{10\varepsilon_1} = \frac{n\,\mathbb{E}\left[\left|U - \frac{1}{n}\right|\right]}{10\varepsilon_1} \leq \frac{1}{10}.$$

Combining the bounds on both terms we get:

$$\Pr[\overline{E^{(1)}}] \leq \frac{2}{10}.$$

Next, we bound the probability of $\overline{E^{(2)}}$.

$$\Pr[\overline{E^{(2)}}] \leq \Pr\left[\left|\sum_{i=1}^n U_i^{(2)} - 1\right| \geq \frac{\varepsilon_2}{10}\right] + \Pr\left[\sum_{i=1}^n \left|U_i^{(2)} - \frac{1}{n}\right| \leq \frac{9\varepsilon_2}{10}\right].$$

Again, note that $U_i^{(2)} \in [a,b]$ and $\mathbb{E}\left[U_i^{(2)}\right] = \frac{1}{n}$. Applying Lemma 32, we obtain

$$\Pr\left[\left|\sum_i U_i^{(2)} - 1\right| \geq \sqrt{10(b-a)}\right] \leq \frac{1}{10},$$

which, since $10(b-a) \leq \frac{\varepsilon_2^2}{100}$, upper bounds the first term in the expression.

Next, we bound the second term. Recall that random variables $U_i^{(2)}$ are independent copies of U'. Since $U' \in [a,b]$ and $a \leq \frac{1}{n} \leq b$, then $0 \leq \left|U' - \frac{1}{n}\right| \leq (b-a)$. Applying Lemma 32, we obtain

$$\Pr\left[\sum_i \left|U_i^{(2)} - \frac{1}{n}\right| \leq n\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right] - \sqrt{10n(b-a)\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right]}\right] \leq \frac{1}{10}.$$

Since $10(b-a) \leq \frac{\varepsilon_2^2}{100} \leq \frac{\varepsilon_2}{100}$ and $\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right] \geq \frac{\varepsilon_2}{n}$, then

$$n\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right] - \sqrt{10n(b-a)\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right]} \geq n\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right] - \frac{n\,\mathbb{E}\left[\left|U' - \frac{1}{n}\right|\right]}{10} \geq \frac{9\varepsilon_2}{10}.$$

Combining the above two equations we get

$$\Pr\left[\sum_i \left|U_i^{(2)} - \frac{1}{n}\right| \leq \frac{9\varepsilon_2}{10}\right] \leq \frac{1}{10}.$$

Combining the bounds on both terms:

$$\Pr[\overline{E^{(2)}}] \leq \frac{2}{10}.$$

Note that the event $E^{(1)}$ implies that

$$\begin{aligned}
\|D^{(1)} - \mathrm{Unif}_n\|_1 &= \sum_{i=1}^{n} \left| \frac{U_i^{(1)}}{\sum_i U_i^{(1)}} - \frac{1}{n} \right| \\
&\leq \sum_{i=1}^{n} \left(\left| \frac{U_i^{(1)}}{\sum_i U_i^{(1)}} - \frac{1}{n \sum_i U_i^{(1)}} \right| + \left| \frac{1}{n} - \frac{1}{n \sum_i U_i^{(1)}} \right| \right) \\
&= \frac{\sum_i |U_i^{(1)} - 1/n|}{\sum_i U_i^{(1)}} + \left| 1 - \frac{1}{\sum_i U_i^{(1)}} \right| \\
&= \frac{\sum_i |U_i^{(1)} - 1/n|}{\sum_i U_i^{(1)}} + \frac{\left| \sum_i U_i^{(1)} - 1 \right|}{\sum_i U_i^{(1)}} \\
&\leq 2 \cdot \frac{\sum_i |U_i^{(1)} - 1/n|}{\sum_i U_i^{(1)}} \\
&\leq 2 \cdot \frac{10\varepsilon_1}{1 - 1/10} \leq \frac{200\varepsilon_1}{9} \leq 25\varepsilon_1. \qquad (31)
\end{aligned}$$

Similarly, event $E^{(2)}$ implies that

$$\begin{aligned}
\|D^{(2)} - \mathrm{Unif}_n\|_1 &= \sum_{i=1}^{n} \left| \frac{U_i^{(2)}}{\sum_i U_i^{(2)}} - \frac{1}{n} \right| \\
&\geq \sum_{i=1}^{n} \left(\left| \frac{U_i^{(2)}}{\sum_i U_i^{(2)}} - \frac{1}{n \sum_i U_i^{(2)}} \right| - \left| \frac{1}{n} - \frac{1}{n \sum_i U_i^{(2)}} \right| \right) \\
&= \frac{\sum_i |U_i^{(2)} - 1/n|}{\sum_i U_i^{(2)}} - \left| 1 - \frac{1}{\sum_i U_i^{(2)}} \right| \\
&= \frac{\sum_i |U_i^{(2)} - 1/n|}{\sum_i U_i^{(2)}} - \frac{\left| \sum_i U_i^{(2)} - 1 \right|}{\sum_i U_i^{(2)}} \\
&\geq \frac{(9\varepsilon_2/10) - (\varepsilon_2/10)}{1 + \varepsilon_2/10} \geq \frac{8\varepsilon_2}{11} \geq \varepsilon_2/2, \qquad (32)
\end{aligned}$$

where we used the triangle inequality and $\varepsilon_2 \leq 1$.

For $j = 1, 2$, let $\mathcal{C}^{(j)}$ denote the distribution of $(C_1^{(j)} \ldots C_n^{(j)})$, and let $\mathcal{P}_{|E^{(j)}}^{(j)}$, $\mathcal{D}_{|E^{(j)}}^{(j)}$ and $\mathcal{C}_{|E^{(j)}}^{(j)}$ denote the distributions of $(U_1^{(j)} \ldots U_n^{(j)})$, $D^{(j)}$, and $(C_1^{(j)} \ldots C_n^{(j)})$, respectively conditioned on the event $E^{(j)}$.

In light of Equations (31) and (32), to prove the lemma it suffices to show no tester using $m/2$ samples from p correctly identifies whether $p = \mathcal{D}_{|E^{(1)}}^{(1)}$ or $p = \mathcal{D}_{|E^{(2)}}^{(2)}$ with probability $\geq 4/5$. To prove by contradiction, suppose there is such a tester $\mathcal{T}$.

Event $E^{(j)}$ implies that $\sum U_i^{(j)} \geq \frac{9}{10}$. Hence, for any given $(U_1^{(j)} \ldots U_n^{(j)}) \sim \mathcal{P}^{(j)}_{|E^{(j)}}$, we have

$$\Pr(N^{(j)} \geq m/2) = \Pr[\mathrm{Poi}(m \sum U_i^{(j)}) \geq m/2] \geq \Pr[\mathrm{Poi}(0.9m) \geq m/2],$$

which is at least 0.95 for m larger than some absolute constant c.

If $N^{(j)} \geq m/2$, then we can simulate $m/2$ samples from $D^{(j)}$ using $(C_1^{(j)} \ldots C_n^{(j)})$ and use the tester $\mathcal{T}$ on these $m/2$ samples. Hence, using this tester we can correctly identify whether $(C_1 \ldots C_n) \sim \mathcal{C}^{(1)}_{|E^{(1)}}$ or $(C_1 \ldots C_n) \sim \mathcal{C}^{(2)}_{|E^{(2)}}$ with probability $\geq 0.95\left(\frac{4}{5}\right) = 0.76$.

Next, we show that the TV distance between the distributions $\mathcal{C}^{(1)}_{|E^{(2)}}$ and $\mathcal{C}^{(2)}_{|E^{(2)}}$ is small.

$$\begin{aligned}
&\mathrm{TV}\Big(\mathcal{C}^{(1)}_{|E^{(1)}}, \mathcal{C}^{(2)}_{|E^{(2)}}\Big) \leq \Pr[\overline{E^{(1)}}] + \mathrm{TV}\Big(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}\Big) + \Pr[\overline{E^{(2)}}] \\
&= \mathrm{TV}\Bigg(\mathop{\mathbb{E}}_{\mathcal{P}^{(1)}}\Big(\mathrm{Poi}(mU_1^{(1)}), \ldots, \mathrm{Poi}(mU_n^{(1)})\Big), \mathop{\mathbb{E}}_{\mathcal{P}^{(2)}}\Big(\mathrm{Poi}(mU_1^{(1)}), \ldots, \mathrm{Poi}(mU_n^{(1)})\Big)\Bigg) + \Pr[\overline{E^{(1)}}] + \Pr[\overline{E^{(2)}}] \\
&\leq n\, \mathrm{TV}\big(\mathbb{E}\,\mathrm{Poi}(mU), \mathbb{E}\,\mathrm{Poi}(mU')\big) + \Pr[\overline{E^{(1)}}] + \Pr[\overline{E^{(2)}}] \\
&\leq n \cdot \frac{1}{20n} + \Pr[\overline{E^{(1)}}] + \Pr[\overline{E^{(2)}}] \leq \frac{9}{20}.
\end{aligned}$$

This implies that for any tester the probability of correctly distinguishing $\mathcal{C}^{(1)}_{|E^{(1)}}$ and $\mathcal{C}^{(2)}_{|E^{(2)}}$ is at most $\frac{1+9/20}{2} = 29/40 = 0.725$, which is a contradiction since $0.725 < 0.76$.

C.5. Proof of Lemma 23

For any finite subset $\mathcal{S}$ of $\subset [-B, A]$, consider the optimization problem

$$\begin{aligned}
\max \mathbb{E}\,|U'| \text{ s.t. } & \mathbb{E}\,|U| \leq \frac{\varepsilon_1}{2} \text{ and} \\
& \mathbb{E}\, U^i = \mathbb{E}\, U'^i, i = 1, \ldots, L, \text{ and} \\
& U, U' \in \mathcal{S}\ ,
\end{aligned} \tag{33}$$

and its dual

$$\begin{aligned}
\min \frac{\varepsilon_1}{2}\alpha + z_1 + z_2 \text{ s.t. } & z_1 + \sum_{i=1}^{L} c_i p_i(x) \geq |x| \text{ for all } x \in \mathcal{S}, \\
& \alpha\,|x| \geq \sum_{i=1}^{L} c_i p_i(x) - z_2 \text{ for all } x \in \mathcal{S}, \text{ and} \\
& \alpha \geq 0\ .
\end{aligned} \tag{34}$$

For a given $\mathcal{S}$, let $P_{\mathcal{S}}$ and $D_{\mathcal{S}}$ be the optimal solution to the primal and dual, respectively. Since $\mathcal{S}$ is finite, the distribution of both U U' is a finite vector of size $\mathcal{S}$, then from the strong duality for linear programming we have $P_{\mathcal{S}} = D_{\mathcal{S}}$.

Let P and D denote the value of optimal solution of (21) and (22), respectively. From the weak duality we have $P \leq D$.

For any $\mathcal{S}$, the corresponding optimization problem (33) can be obtained by imposing the constraints $\Pr[U \in [-B, A] \setminus \mathcal{S}] = \Pr[U' \in [-B, A] \setminus \mathcal{S}] = 0$ in (21). Since upon imposing the additional constrains, the value of the optimal solution in (21) would only decrease, hence $P_{\mathcal{S}} \leq P$. This implies for all finite subset $\mathcal{S}$ of $\subset [-B, A]$, the following holds $P \geq P_{\mathcal{S}} = D_{\mathcal{S}}$.

Let $\mathcal{S}_\delta = \{x : x = -B + k\delta \text{ for } k \in \{0, 1, \ldots, \lfloor \frac{A+B}{\delta} \rfloor\}\}$. Observe that for all $\delta > 0$, $\mathcal{S}_\delta$ is a finite subset of $[-B, A]$. Taking the supremum over $\mathcal{S}_\delta$ as $\delta \to 0$,

$$P \geq \sup_{\delta > 0} D_{\mathcal{S}_\delta}.$$

Using the continuity of functions x^i and $|x|$ and elementary real analysis it can be verified that

$$\sup_{\delta > 0} D_{\mathcal{S}_\delta} = D.$$

Hence, we get $P \geq D$. Combining this with $P \leq D$ proves the lemma. □

Appendix D. Instance-optimal tolerant testing

In this appendix, we establish our "instance-optimal" tolerant identity bounds (Theorem 3); that is, sample complexity bounds parameterized by the reference distribution q itself, instead of the domain size n. We do so by establishing separately the lower bound (Theorem 38) and upper bound (Theorem 39) parts of the statement, in subsection D.1 and subsection D.2.[3]

In order to formally state our results, a few definitions will be useful. For any distribution q over a set $[n]$ and any subset $S \subseteq [n]$, let $\|q_S\|_\infty := \max_{i \in S} q_i$, and $\rho_{q,S} := \left\lfloor \frac{q(S)}{2\|q_S\|_\infty} \right\rfloor$, where as usual $q(S) = \sum_{i \in S} q_i$. Moreover, for any $x \geq 0$, let q_{-x} denote the vector obtained by iteratively removing the smallest entries from q and stopping just before the sum of the removed elements exceed x. Finally, recall that for any integer $t \geq 1$, Unif_t denotes the uniform distribution over $[t]$.

D.1. Lower bound

Given a reference distribution q, $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ and $\delta > 0$, let $\mathrm{SC}(q, \varepsilon_1, \varepsilon_2, \delta)$ denote the minimum number of samples (in the Poissonized sampling model) any tester requires from an unknown distribution p to correctly distinguish between $\|p - q\|_1 \leq \varepsilon_1$ and $\|p - q\|_1 \geq \varepsilon_2$ with probability at least $1 - \delta$.

Our main tool will be the following theorem relating the lower bound of testing uniform distributions to the lower bound of testing for general q.

Theorem 34 *For any distribution q over $[n]$, subset $S \subseteq [n]$ such that $\rho_{q,S} \geq 1$, $0 \leq \varepsilon_1 < \varepsilon_2$ and $\delta > 0$,*

$$\mathrm{SC}(q, \varepsilon_1, \varepsilon_2, \delta) \geq \frac{4}{q(S)} \cdot \mathrm{SC}(\mathrm{Unif}_{\rho_{q,S}}, \tfrac{4\varepsilon_1}{q(S)}, \tfrac{4\varepsilon_2}{q(S)}, \delta).$$

Proof For any $S \subseteq [n]$ and any distribution p over $[\rho_{q,S}]$, we derive a distribution p^{new} over $[n]$ such that $\|p^{\text{new}} - q\|_1 = \frac{q(S)}{4}\|p - \mathrm{Unif}_{\rho_{q,S}}\|_1$ and, for any $m > 0$, $\mathrm{Poi}(m)$ samples from p can be used to generate $\mathrm{Poi}\left(4m/q(S)\right)$ samples from p^{new}, with the knowledge of just q and not of p and p^{new}. Then the statement of the theorem follows, since to distinguish $\|p - \mathrm{Unif}_{\rho_{q,S}}\|_1 \leq \frac{4\varepsilon_1}{q(S)}$

[3] As mentioned earlier, we slightly abuse the $\tilde{O}$ and $\tilde{\Omega}$ notation in those two statements to also hide logarithmic factors in n, not just in the argument.

and $\|p - \mathrm{Unif}_{\rho_{q,S}}\|_1 \le \frac{4\varepsilon_2}{q(S)}$ one can use samples from p to generate samples from p^{new} and test $\|p^{\text{new}} - q\|_1 \le \varepsilon_1$ vs $\|p^{\text{new}} - q\|_1 \ge \varepsilon_2$ instead.

The rest of the proof focuses on obtaining such a distribution p^{new}.

Fix any S is a subset of $[n]$ such that $\rho_{q,S} \ge 1$, and let $\|q_S\|_\infty = \max_{i\in S} q_i$. Consider a partition of S into $S_1, S_2, \dots, S_\ell$ (for some $\ell \ge 1$) such that $q(S_j) \in [\|q_S\|_\infty, 2\|q_S\|_\infty)$ for every $j \in [\ell]$. Such partition exists, and can be obtained by a greedy construction. Since the mass of each S_j is less than $2\|q_S\|_\infty$, we also have that $\ell > \frac{q(S)}{2\|q_S\|_\infty} \ge \rho_{q,S}$.

Given a distribution p on $[\rho_{q,S}]$, we define p^{new} as follows. For every $1 \le j \le \rho_{q,S}$ (i.e., the first $\rho_{q,S}$ subsets), each element $i \in S_j$ is given the probability

$$p_i^{\text{new}} = q_i + q_i \cdot \frac{q(S)}{4q(S_j)}\left(p_j - \frac{1}{\rho_{q,S}}\right) \tag{35}$$

while every $i \in \bigcup_{j>\rho_{q,S}} S_j$ is assigned the probability $p_i^{\text{new}} = q_i$. Next we show that p^{new} is indeed a distribution, and can be sampled given samples from p and knowledge of q only.

- For $i \notin \bigcup_{j\le\rho_{q,S}} S_j$, since $p_i^{\text{new}} = q_i$, we have $p_i^{\text{new}} \ge 0$. Moreover, the count $\mathrm{Poi}(mq_i)$ can clearly be generated with the knowledge of q, m only.

- For $j \le \rho_{q,S}$ and any element $i \in S_j$, note that

  $$p_i^{\text{new}} = q_i + q_i\frac{q(S)}{4q(S_j)}\left(p_j - \frac{1}{\rho_{q,S}}\right) \ge q_i - q_i\frac{q(S)}{4q(S_j)}\cdot\frac{1}{\rho_{q,S}} \ge q_i - q_i\frac{q(S)}{4\|q_S\|_\infty}\cdot\frac{1}{\rho_{q,S}} \ge q_i - q_i\frac{\rho_{q,S}+1}{2\rho_{q,S}} \ge 0.$$

  Using the standard properties of Poisson processes it is easy to see that for any $j \le \rho_{q,S}$, a sample from $\mathrm{Poi}(mp_j)$ and the knowledge of q suffice to generate $\mathrm{Poi}(\frac{4m}{q(S)}p_i^{\text{new}})$ samples for each $i \in S_j$.

- Finally,

  $$\sum_{i\in[n]} p_i^{\text{new}} = \sum_{i\in[n]} q_i + \sum_{j\le\rho_{q,S}} q(S_j)\frac{q(S)}{4q(S_j)}\left(p_j - \frac{1}{\rho_{q,S}}\right) = 1 + \frac{q(S)}{4}\left(\sum_{j\le\rho_{q,S}} p_j - 1\right) = 1.$$

  This shows that p^{new} is indeed a distribution.

To complete the proof, it only remains to relate the ℓ_1 distances, which we do now.

$$\sum_{i\in[n]} |p_i^{\text{new}} - q_i| = \sum_{j\le\rho_{q,S}}\sum_{i\in S_j} q_i\frac{q(S)}{4q(S_j)}\left|p_j - \frac{1}{\rho_{q,S}}\right| = \frac{q(S)}{4}\sum_{j\le\rho_{q,S}}\left|p_j - \frac{1}{\rho_{q,S}}\right| = \frac{q(S)}{4}\|p - \mathrm{Unif}_{\rho_{q,S}}\|_1\,,$$

as claimed. ■

The lower bound from Theorem 7 implies that for any subset S such that $\rho_{q,S} \ge 2$ and $0 \le \frac{4\varepsilon_1}{q(S)} < \frac{4\varepsilon_2}{q(S)} \le c$ (for some universal constant $c > 0$),

$$\mathrm{SC}(\mathrm{Unif}_{\rho_{q,S}}, \frac{4\varepsilon_1}{q(S)}, \frac{4\varepsilon_2}{q(S)}, 4/5) = \Omega\left(\frac{q(S)\cdot\rho_{q,S}}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{q(S)^2\rho_{q,S}}{\log n}\left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2\right).$$

Combining this bound with the above theorem and using the observation that if $\rho_{q,S} \le 2$ then $\rho_{q,S} - 2 \le 0$ and $\rho_{q,S}\cdot q(S) \le 0$, we get:

Corollary 35 *For any distribution q over $[n]$, $0 \leq \varepsilon_1 < \varepsilon_2 < 1$, and some universal constant $c > 0$,*

$$\mathrm{SC}(q, \varepsilon_1, \varepsilon_2, 4/5) \geq \Omega\left(\max_{S \subseteq [n] : q(S) \geq 4\varepsilon_2/c,} \left((\rho_{q,S} - 2) \cdot \frac{1}{\log n} \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right) + (\rho_{q,S} - 2) \cdot \frac{1}{\log n} \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right)^2 \right) \right).$$

We would like to relate this bound, which involves a maximum over subsets S and the quantity $\rho_{q,S}$, to a more interpretable expression involving the 0- and $1/2$-quasinorms of q, as stated in Theorem 3. Our next two lemmas will allow us to do so.

Lemma 36 *For any $x \in (0, 1)$ such that $\|q_{-x}\|_0 > 1$, there exists some $i^* \in [n]$ such that for $A := \{j : q_j \leq q_i^*\}$ the following holds: (i) $\max_{j \in A} q_j = q_{i^*}$, (ii) $q(A) \geq x$ and (iii) $\frac{q(A)}{q_{i^*}} \geq \frac{\|q_{-x}\|_0 - 1}{\ln(1/x)}$.*

Proof Without loss of generality, we can assume that the distribution $q = (q_1, q_2, ..., q_n)$ is non-increasing, that is that $q_1 \geq q_2 \geq .... \geq q_n$. This in particular implies that $\|q_{-x}\|_0 = \max\{i : \sum_{j \geq i} q_j \geq x\}$.

Note that for any $i^* \in [n]$ property (i) holds trivially, and property (ii) holds for any $i^* \leq \|q_{-x}\|_0$, as $q(A) = \sum_{j \in A} q_j = \sum_{j : q_j \leq q_{i^*}} q_j \geq \sum_{j : j \geq i^*} q_j \geq \sum_{j : j \geq \|q_{-x}\|_0} q_j \geq x$.

To complete the proof, it thus suffices to establish (iii) for some $i^* \leq \|q_{-x}\|_0$; that is, to find $i^* \leq \|q_{-x}\|_0$ such that

$$\frac{\sum_{j \in A} q_j}{q_{i^*}} \geq \frac{\sum_{j \geq i^*} q_j}{q_{i^*}} \geq \frac{\|q_{-x}\|_0 - 1}{\ln(1/x)}.$$

Since if $\frac{\|q_{-x}\|_0 - 1}{\ln(1/x)} \leq 1$ this trivially holds for every $i^* \leq \|q_{-x}\|_0$, in what follows we assume $\frac{\|q_{-x}\|_0 - 1}{\ln(1/x)} > 1$.

Suppose by contradiction that for every $i \leq \|q_{-x}\|_0$, we have $\frac{\sum_{j \geq i} q_j}{q_i} \leq \frac{\|q_{-x}\|_0 - 1}{\ln(1/x)}$; equivalently, that $q_i \geq (\sum_{j \geq i} q_j) \frac{\ln(1/x)}{\|q_{-x}\|_0 - 1}$. Hence, $\sum_{j \geq i+1} q_j = (\sum_{j \geq i} q_j) - q_i \leq (\sum_{j \geq i} q_j) \Big(1 - \frac{\ln(1/x)}{\|q_{-x}\|_0 - 1}\Big)$. By induction, this gives

$$\sum_{j \geq \|q_{-x}\|_0} q_j \leq \left(\sum_{j \geq 1} q_j \right) \left(1 - \frac{\ln(1/x)}{\|q_{-x}\|_0 - 1} \right)^{\|q_{-x}\|_0 - 1} = \left(1 - \frac{\ln(1/x)}{\|q_{-x}\|_0 - 1} \right)^{\|q_{-x}\|_0 - 1} < e^{-\ln(1/x)} = x\,,$$

where we used that $0 \leq 1 - u < e^{-u}$ for $u \in (0, 1)$. But, by definition $\sum_{j \geq \|q_{-x}\|_0} q_j \geq x$: this is a contradiction, concluding the proof. ■

Lemma 37 *For any distribution q over $[n]$ and $x \in (0, 1)$, $\max_{S \subseteq [n] : q(S) \geq x} \rho_{q,S} \geq \frac{\|q_{-x}\|_{1/2}}{(\log(n/x)+1)^2} - 4$.*

Proof Let $D = \max\{S : q(S) \leq x\}$ be a largest subset that has mass $\leq x$ under q. From the definition of D it is not hard to see that we can choose D such that $\min_{i \in [n] \setminus D} q_i \geq \max_{i \in D} q_i$, and

$$x < \sum_{i \in D} q_i + \min_{i \in [n] \setminus D} q_i \leq (D + 1) \min_{i \in [n] \setminus D} q_i \leq n \min_{i \in [n] \setminus D} q_i,$$

therefore, $\min_{i \in [n] \setminus D} q_i > \frac{x}{n}$.

Next, we perform a "bucketing" of the remaining elements; that is, we partition $[n]\setminus D$ in subsets so that the probability assigned by q to any two elements in the same subset differ by at most a factor 2. Let $\ell = \lfloor \log(\frac{n}{x}) \rfloor + 1$ and for $j \in [\ell]$, let

$$D_j := \left\{ i \in [n] \setminus D : q_i \in \left(\frac{1}{2^j}, \frac{1}{2^{j-1}} \right] \right\}.$$

We can write

$$\|q_{-x}\|_{1/2} = \left(\sum_{j\in[\ell]} \sum_{i\in D_j} q_i^{1/2} \right)^2 \leq \ell^2 \max_{j\in[\ell]} \left(\sum_{i\in D_j} q_i^{1/2} \right)^2 \leq \ell^2 \max_{j\in[\ell]} |D_j| \cdot q(D_j), \tag{36}$$

the last inequality being Cauchy–Schwarz. Let $j^* \in [\ell]$ be the index maximizing the term on the left, and choose $S = D_{j^*} \cup D$. Since D_{j^*} is non-empty, from the definition of D, we have $q(S) \geq x$. Further,

$$\rho_{q,S} = \left\lfloor \frac{q(S)}{2\|q_S\|_\infty} \right\rfloor = \left\lfloor \frac{\sum_{i\in D_{j^*}\cup D} q_i}{2\max_{i\in D_{j^*}\cup D} q_i} \right\rfloor \geq \left\lfloor \frac{\sum_{i\in D_{j^*}} q_i}{2\max_{i\in D_{j^*}} q_i} \right\rfloor \geq \left\lfloor \frac{|D_{j^*}| \cdot 2^{-j^*}}{2 \cdot 2^{-j^*+1}} \right\rfloor \geq \frac{|D_{j^*}|}{4} - 1. \tag{37}$$

Putting together (36) and (37), we get

$$\frac{\|q_{-x}\|_{1/2}}{\ell^2} \leq |D_{j^*}| \cdot q(D_{j^*}) \leq |D_{j^*}| \cdot q(S) \leq (\rho_{q,S} + 4)(q(S)) \leq \rho_{q,S} \cdot q(S) + 4$$

which concludes the proof. ■

Combining Corollary 35 and the above two lemmas for $x = 4\varepsilon_2/c$, for some universal constant $c > 0$, any distribution q over $[n]$, $0 \leq \varepsilon_1 < \varepsilon_2 < c/4$,

$$\mathrm{SC}(q, \varepsilon_1, \varepsilon_2, 4/5) \geq \Omega\left(\left(\frac{\|q_{-4\varepsilon_2/c}\|_0}{\ln(c/4\varepsilon_2)} - 3 \right) \cdot \frac{1}{\log n} \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right) + \left(\frac{||q_{-4\varepsilon_2/c}||_{1/2}}{\log^2(nc/4\varepsilon_2)} - 6 \right) \cdot \frac{1}{\log n} \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right)^2 \right).$$

By combining the above lower bound with previously known lower bound $\Omega(\frac{||q_{-c'\varepsilon_2}||_{2/3}-1}{\varepsilon_2^2})$ for non-tolerant identity testing from Valiant and Valiant (2014), where $c' > 0$ is an absolute constant, we obtain:

Theorem 38 *For any distribution q over $[n]$, $0 \leq \varepsilon_1 < \varepsilon_2 < c/4$, for some universal constant $c > 0$,*

$$\mathrm{SC}(q, \varepsilon_1, \varepsilon_2, 4/5) \geq \tilde{\Omega}\left(||q_{-4\varepsilon_2/c}||_0 \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right) + ||q_{-4\varepsilon_2/c}||_{1/2} \left(\frac{\varepsilon_1}{\varepsilon_2^2} \right)^2 + \frac{||q_{-4\varepsilon_2/c}||_{2/3}}{\varepsilon_2^2} \right) - \mathcal{O}\left(\frac{1}{\varepsilon_2^2} \right).$$

This establishes the lower bound part of Theorem 3.

D.2. Upper bound

The proof of our instance-optimal upper bound follows the same outline as (Diakonikolas and Kane, 2016, Proposition 2.12), yet the extension to tolerant testing requires a significantly more detailed argument.

Theorem 39 (Identity testing) *Let q be a known reference distribution and p be an . There is a computationally efficient algorithm with the following guarantee. Given a known reference distribution q over $[n]$, as well as parameters $\varepsilon_1, \varepsilon_2$ such that $0 \leq \varepsilon_2 \leq 1$ and $0 \leq \varepsilon_1 \leq c\frac{\varepsilon_2}{\log(n/\varepsilon_2)}$ (where $c > 0$ is an absolute constant), the algorithm takes*

$$\tilde{O}\left(\|q_{-\varepsilon_2/20}\|_{\frac{1}{2}} \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \|q_{-\varepsilon_2/20}\|_0 \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{\|q_{-\varepsilon_2/20}\|_{\frac{2}{3}}}{\varepsilon_2^2} \right).$$

samples from an unknown distribution p over $[n]$ and distinguishes between $\|p-q\|_1 \leq \varepsilon_1$ and $\|p-q\|_1 \geq \varepsilon_2$ with probability at least $4/5$.

Proof Let $D = \operatorname{argmax}\{S : \sum_{i\in S} q_i \leq \varepsilon_2/20\}$ be a largest subset that has mass $\leq \varepsilon_2/20$. From the definition of D it is easy to see that we can choose D such that $\min_{i\in[n]\setminus D} q_i \geq \max_{i\in D} q_i$, and thus

$$\frac{\varepsilon_2}{20} < \sum_{i\in D} q_i + \min_{i\in[n]\setminus D} q_i \leq (D+1) \min_{i\in[n]\setminus D} q_i \leq n \min_{i\in[n]\setminus D} q_i,$$

which implies that $\min_{i\in[n]\setminus D} q_i > \frac{\varepsilon_2}{20n}$. Moreover, given the full description of q, this set D can be efficiently computed. We then (as in the lower bound section) "bucket" the remaining elements $[n] \setminus D$ into disjoint subsets so that the probability assigned by q to any two elements in the same subset differ by at most a factor 2. That is, for $\ell := \lfloor \log(\frac{20n}{\varepsilon_2}) \rfloor + 1$ and $j \in [\ell]$, we let

$$D_j := \left\{ i \in [n] \setminus D : q_i \in \left(\frac{1}{2^j}, \frac{1}{2^{j-1}} \right] \right\}. \tag{38}$$

We denote by p^j and q^j the conditional distributions on D_j induced by p and q, respectively. With this in hand, we get the following:

Claim 40 *If $\|p-q\|_1 \leq \varepsilon_1$, then all three conditions below hold simultaneously:*

1. *$p(D) \leq \varepsilon_1 + \frac{\varepsilon_2}{20}$ and*
2. *for every $j \in [\ell]$, $|p(D_j) - q(D_j)| \leq \varepsilon_1$, and*
3. *for every $j \in [\ell]$, $\|p^j - q^j\|_1 \leq \frac{2\varepsilon_1}{q(D_j)}$.*

Proof We prove the claim by showing that if any of the three conditions fails to hold then we must have $\|p-q\|_1 \geq \varepsilon_1$. Note that, by the triangle inequality,

$$\|p-q\|_1 = \sum_{i\in D} |p_i - q_i| + \sum_{j=1}^{\ell} \sum_{i\in D_j} |p_i - q_i| \geq |p(D) - q(D)| + \sum_{j=1}^{\ell} |p(D_j) - q(D_j)|.$$

Recalling that $q(D) \leq \frac{\varepsilon_2}{20}$, it is easy to see that if either of the first two conditions fails to hold, the above inequality implies that $\|p - q\|_1 \geq \varepsilon_1$. Turning to the third condition, we can write

$$\begin{aligned}\|p^j - q^j\|_1 &= \sum_{i \in D_j} \left| \frac{p_i}{p(D_j)} - \frac{q_i}{q(D_j)} \right| \leq \sum_{i \in D_j} \left| \frac{p_i}{q(D_j)} - \frac{q_i}{q(D_j)} \right| + \sum_{i \in D_j} \left| \frac{p_i}{p(D_j)} - \frac{p_i}{q(D_j)} \right| \\ &= \frac{\sum_{i \in D_j} |p_i - q_i| + |p(D_j) - q(D_j)|}{q(D_j)} \leq \frac{2\sum_{i \in D_j} |p_i - q_i|}{q(D_j)} \leq \frac{2\|p - q\|_1}{q(D_j)} ,\end{aligned}$$

which shows that if the third item fails to hold for some j, then $\|p - q\|_1 > \varepsilon_1$. ■

The next claim then provides a qualitatively converse statement.

Claim 41 *Suppose that* p *satisfies all three conditions below:*

1. $p(D) \leq \frac{\varepsilon_2}{5}$,
2. *for every* $j \in [\ell]$, $|p(D_j) - q(D_j)| \leq \frac{\varepsilon_2}{10\ell}$, *and*
3. *for every* $j \in [\ell]$ *such that* $q(D_j) \geq \frac{\varepsilon_2}{5\ell}$, $\|p^j - q^j\|_1 \leq \frac{\varepsilon_2}{5\ell q(D_j)}$.

Then, we have $\|p - q\|_1 \leq \varepsilon_2$.

Proof Suppose that the three conditions hold. By the triangle inequality, we have

$$\begin{aligned}\|p - q\|_1 &= \sum_{i \in D} |p_i - q_i| + \sum_{j=1}^{\ell} \sum_{i \in D_j} |p_i - q_i| \\ &\leq p(D) + q(D) + \sum_{j=1}^{\ell} \sum_{i \in D_j} q(D_j) \left| \frac{p_i}{q(D_j)} - \frac{q_i}{q(D_j)} \right| \\ &\leq p(D) + \frac{\varepsilon_2}{20} + \sum_{j=1}^{\ell} \sum_{i \in D_j} q(D_j) \left| \frac{p_i}{p(D_j)} - \frac{q_i}{q(D_j)} \right| + \sum_{j=1}^{\ell} \sum_{i \in D_j} q(D_j) \left| \frac{p_i}{p(D_j)} - \frac{p_i}{q(D_j)} \right| \\ &\leq p(D) + \frac{\varepsilon_2}{20} + \sum_{j=1}^{\ell} q(D_j) \|p^j - q^j\|_1 + \sum_{j=1}^{\ell} |p(D_j) - q(D_j)| \\ &\leq \frac{\varepsilon_2}{5} + \frac{\varepsilon_2}{20} + \left(\ell \cdot \frac{\varepsilon_2}{5\ell} \cdot 2 + \sum_{j=1}^{\ell} q(D_j) \cdot \frac{\varepsilon_2}{5\ell q(D_j)} \right) + \ell \cdot \frac{\varepsilon_2}{10\ell} < \varepsilon_2 ,\end{aligned}$$

where we used the three conditions for the second-to-last inequality. ■

Given the above claims, we can describe our testing algorithm. First, the algorithm computes the set D, the value ℓ, and the bucketing of $[n] \setminus D$ into $D_1, \dots, D_\ell$. Then, it runs a total of (at most) $2\ell + 1$ sub-tests, which we will detail momentarily:

(1) Distinguish $p(D) < \frac{\varepsilon_2}{20} + \varepsilon_1$ (accept) from $p(D) \geq \frac{\varepsilon_2}{5}$ (reject),

(2) For every $j \in [\ell]$, distinguish $|p(D_j) - q(D_j)| \leq \varepsilon_1$ (accept) from $|p(D_j) - q(D_j)| \geq \frac{\varepsilon_2}{10\ell}$ (reject), and

(3) For every $j \in [\ell]$ such that $q(D_j) \geq \frac{\varepsilon_2}{5\ell}$, distinguish $\|p^j - q^j\|_1 \leq \frac{2\varepsilon_1}{\ell q(D_j)}$ (accept) from $\|p^j - q^j\|_1 \geq \frac{\varepsilon_2}{5\ell q(D_j)}$ (reject).

If all the above testers accept, the overall tester accepts (i.e., outputs $\|p - q\|_1 \leq \varepsilon_1$); otherwise, it rejects (i.e., outputs $\|p - q\|_1 \geq \varepsilon_2$).

From Claims 40 and 41, it is not hard to see that if $\varepsilon_1 \leq \varepsilon_2/(40\ell)$ and all the above testers give correct outputs with probability at least $1 - \frac{1}{5(2\ell+1)}$ each, then we correctly distinguish $\|p-q\|_1 \leq \varepsilon_1$ and $\|p - q\|_1 \geq \varepsilon_2$ with probability at least $1 - 1/5 = 4/5$ (by a union bound). We now proceed to describe how those tests are implemented.

- Using $\mathcal{O}\Big(\frac{\log(1/\ell)}{(\varepsilon_2/\ell)^2}\Big) = O\Big(\frac{\ell^2 log(1/\ell)}{\varepsilon_2^2}\Big)$ samples from p one can estimate $p(D)$ and $p(D_j)$ for every $j \in [\ell]$ to an additive $\varepsilon_2/(20\ell)$ with probability at least $1 - \frac{1}{5(2\ell+1)}$ each, which gives us the testers for (1) and (2).

- Theorem 5 provides a tester that, for any fixed j, distinguishes between $\|p^j - q^j\|_1 \leq \frac{2\varepsilon_1}{\ell q(D_j)}$ and $\|p^j - q^j\|_1 \geq \frac{\varepsilon_2}{5\ell q(D_j)}$ with probability of success $\geq 4/5$ and uses

$$\mathcal{O}\left(|D_j|\Big(\frac{\varepsilon_1(\ell \cdot q(D_j))}{\varepsilon_2^2}\Big)^2 + |D_j|\Big(\frac{\varepsilon_1(\ell q(D_j)}{\varepsilon_2^2}\Big) + \frac{(\ell \cdot q(D_j))^2\sqrt{|D_j|}}{\varepsilon_2^2}\right)$$

  samples from p^j. By standard amplification arguments one can achieve a probability of success of $1 - \frac{1}{10(2\ell+1)}$ at the cost of a multiplicative $\mathcal{O}(\log(1/\ell))$ factor in the sample complexity.

  To use this in order to obtain the tests required for (3), note that for any $j \in [\ell]$ such that $q(D_j) \geq \frac{\varepsilon_2}{5\ell}$ if $|p(D_j) - q(D_j)| \geq \frac{\varepsilon_2}{10\ell}$ then the corresponding test from (2) already outputs reject with high probability; so we can assume that $p(D_j) \geq q(D_j) - \frac{\varepsilon_2}{10\ell} \geq q(D_j)/2$. In this case for any $m > 0$, using m samples from p, we can get $\Omega(mp(D_j)/\log \ell) = \Omega(mq(D_j)/\log \ell)$ samples from p^j with probability at least $1 - 1/(10\ell^2)$. Note that we can use the same overall set of m samples from p to obtain our $m_j = \Omega(mq(D_j)/\log \ell)$ samples from every p^j, $j \in [\ell]$.

  This gives us the testing algorithms for (3).

Combining these bounds, we get the following upper bound on the sample complexity:

$$\begin{aligned}
&\mathcal{O}\Big(\frac{\ell^2 \log \ell}{\varepsilon_2^2}\Big) + \max_{j \in \ell} \frac{\log \ell}{q(D_j)} \cdot \mathcal{O}\left(|D_j|\Big(\frac{\varepsilon_1(\ell \cdot q(D_j))}{\varepsilon_2^2}\Big)^2 + |D_j|\Big(\frac{\varepsilon_1(\ell \cdot q(D_j))}{\varepsilon_2^2}\Big) + \frac{(\ell \cdot q(D_j))^2\sqrt{|D_j|}}{\varepsilon_2^2}\right) \\
&= \tilde{\mathcal{O}}\Big(\frac{1}{\varepsilon_2^2}\Big)) + \max_{j \in \ell} \tilde{\mathcal{O}}\left(|D_j|q(D_j)\Big(\frac{\varepsilon_1}{\varepsilon_2^2}\Big)^2 + |D_j|\Big(\frac{\varepsilon_1}{\varepsilon_2^2}\Big) + \frac{q(D_j)\sqrt{|D_j|}}{\varepsilon_2^2}\right) \\
&= \tilde{\mathcal{O}}\Big(\frac{1}{\varepsilon_2^2}\Big)) + \max_{j \in \ell} \tilde{\mathcal{O}}\left(|D_j|^2 2^{-j}\Big(\frac{\varepsilon_1}{\varepsilon_2^2}\Big)^2 + |D_j|\Big(\frac{\varepsilon_1}{\varepsilon_2^2}\Big) + \frac{2^{-j}(|D_j|)^{3/2}}{\varepsilon_2^2}\right),
\end{aligned}$$

where the last line uses the definition of D_j in (38) (the “bucketing”) to relate $q(D_j)$ to $|D_j|$.

To conclude, we observe that

$$\begin{aligned}
\|q_{-\varepsilon_2/20}\|_0 &= n - |D| \geq \max_{j\in[\ell]} |D_j| \\
\|q_{-\varepsilon_2/20}\|_{\frac{1}{2}} &= \big(\sum_{j\in[\ell]}\sum_{i\in D_j} q_i^{1/2}\big)^2 \geq \max_{j\in[\ell]}(|D_j|2^{-j/2})^2. \\
\|q_{-\varepsilon_2/20}\|_{\frac{2}{3}} &= \big(\sum_{j\in[\ell]}\sum_{i\in D_j} q_i^{2/3}\big)^{3/2} \geq \max_{j\in[\ell]}(|D_j|2^{-2j/3})^{3/2}.
\end{aligned}$$

Combining the above four equations, and using $\|q_{-\varepsilon_2/20}\|_{\frac{2}{3}} = \Omega(1)$, we get the following upper bound on the sample complexity:

$$\tilde{\mathcal{O}}\left(\|q_{-\varepsilon_2/20}\|_{\frac{1}{2}} \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right)^2 + \|q_{-\varepsilon_2/20}\|_0 \left(\frac{\varepsilon_1}{\varepsilon_2^2}\right) + \frac{\|q_{-\varepsilon_2/20}\|_{\frac{2}{3}}}{\varepsilon_2^2} \right).$$

This concludes the proof of the theorem. ■