Bessel Smoothing and Multi-Distribution Property Estimation

Yi Hao[§] YIH179@ENG.UCSD.EDU

Ping Li PINGLI98@GMAIL.COM

Cognitive Computing Lab, Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider a basic problem in statistical learning: estimating properties of multiple discrete distributions. Denoting by Δ_k the standard simplex over $[k] := \{0, 1, \ldots, k\}$, a property of d distributions is a mapping from Δ_k^d to $\mathbb{R}$. These properties include well-known distribution characteristics such as Shannon entropy and support size ($d = 1$), and many important divergence measures between distributions ($d = 2$). The primary problem being considered is to learn the property value of an *unknown* d-tuple of distributions from its sample. The study of such problems dates back to the works of Good (1953); Carlton (1969); Efron and Thisted (1976); Thisted and Efron (1987), and has been pushed forward steadily during the past decades. Surprisingly, before our work, the general landscape of this fundamental learning problem was insufficiently understood, and nearly all the existing results are for the special case $d \leq 2$.

Our first main result provides a near-linear-time computable algorithm that, given independent samples from any collection of distributions and for a broad class of multi-distribution properties, learns the property as well as the empirical plug-in estimator that uses samples with logarithmic-factor larger sizes. As a corollary of this, for any $\varepsilon > 0$ and fixed $d \in \mathbb{Z}^+$, a d-distribution property over $[k]$ that is Lipschitz and additively separable can be learned to an accuracy of ε using a sample of size $\mathcal{O}(k/(\varepsilon^3\sqrt{\log k}))$, with high probability. Our second result addresses a closely related problem – tolerant independence testing: One receives samples from the unknown joint and marginal distributions, and attempts to infer the ℓ_1 distance between the joint distribution and the product distribution of the marginals. We show that this testing problem also admits a sample complexity sub-linear in the alphabet sizes, demonstrating the broad applicability of our approach.

Keywords: Probability Distribution; Property/Functional Estimation; Maximum Likelihood

1. Introduction

Properties of distributions play a fundamental role in statistics, information theory, and machine learning (Good, 1953; Chow and Liu, 1968; McNeil, 1973; Efron and Thisted, 1976; Chao, 1984; Thisted and Efron, 1987; Chao and Lee, 1992; Haas et al., 1995; Mainen and Sejnowski, 1995; van Steveninck et al., 1997; Kroes et al., 1999; Batu et al., 2000, 2001; Gerstner and Kistler, 2002; Mao and Lindsay, 2007; Ionita-Laza et al., 2009; Ron, 2010; Colwell et al., 2012; Cover and Thomas, 2012; Quinn et al., 2013; Chao and Chiu, 2014; Bresler, 2015). Notable examples include the Shannon entropy and support size of a single distribution, and KL divergence and ℓ_1 distance between a pair. Modern data science applications, ranging from inferring the number of unseen species in a

[§] Yi Hao is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of California, San Diego. Yi Hao's work was conducted while he was a summer intern at Baidu Research.

population (Good, 1953; Chao, 1984; Smith and van Belle, 1984; Chao, 2004) to constructing tree-structured models for graphs and images (Chow and Liu, 1968; Quinn et al., 2013; Bresler, 2015), often call for estimation of such properties of unknown distributions.

The study of distribution property estimation dates back several decades to the works of Good (1953); Carlton (1969); Efron and Thisted (1976) and has steadily grown over the years (Paninski, 2003, 2004; Li and Zhang, 2011; Valiant and Valiant, 2011a,b; Jiao et al., 2015; Acharya et al., 2016; Orlitsky et al., 2016; Valiant and Valiant, 2016; Wu and Yang, 2016; Acharya et al., 2017; Hao et al., 2018; Hao and Orlitsky, 2019a,b,c, 2020). A widely used estimation scheme is to plug the empirical distributions into the property. For example, if one wants to infer an unknown distribution's entropy, compute the entropy of its sample empirical distribution. Standard results from statistics (Van der Vaart, 2000) show that this plug-in approach is essentially optimal when the sample size is enormous compared with the underlying alphabet. Nevertheless, modern learning applications often concern high-dimensional data, and usually, the problem's dimension is comparable to or even much larger than the data size. The desire to go beyond the classical large-sample analysis and design algorithms that perform well in such data-sparse regimes has led to the recent advances in the field. For several properties, including the four mentioned above, optimal estimators that are more efficient than the empirical ones have been discovered (Valiant and Valiant, 2011a; Jiao et al., 2015; Orlitsky et al., 2016; Acharya et al., 2016; Wu and Yang, 2016; Acharya et al., 2017; Jiao et al., 2018; Hao et al., 2018; Wu and Yang, 2019; Hao and Orlitsky, 2019a,b,c).

Despite years of research, nearly all existing works focus on cases involving only one or two distributions, and it is often nontrivial to extend their techniques and analysis to the multi-distribution case (see Section 5). However, as shown by our discussion in Section 2, a variety of real-world learning applications require estimating properties of multiple distributions. This gap between theory and practice has become the primary motivation for the present work.

We study the fundamental problem of estimating a general multi-distribution property of unknown discrete distributions from independent samples. Our aim is to emulate the performance of the empirical estimator having access to samples of larger sizes. We show that for a broad class of properties, regardless of the underlying alphabet sizes and for every distribution tuple, it is always possible to *amplify* the sample sizes by logarithmic factors. As an implication of this result, for many important properties considered in the paper, our approach yields the first estimator whose sample complexity is sub-linear in the possibly unknown alphabet sizes. Equally importantly, nearly all the proposed algorithms are near-linear-time computable. These advances enable us to have the first glimpse of the general landscape of distribution property estimation in high dimensions.

Paper outline Section 2 provides six important properties arising in applications and covered by our approach. Section 3 presents major theorems and corollaries, addressing both the classical empirical estimator and our sub-linear sample-complexity estimator. In Section 4, we discuss our main technique and its connection to the prior work (Hao et al., 2018), show our theoretical contributions, and provide the outline of proofs and explicit form of our algorithm. Section 5 reviews major prior results, makes comparisons, and illustrates why the corresponding methods do not easily adapt to our setting. For an outline of the technical part of the paper – the appendices, see Appendix A.

2. Six Examples of Multi-Distribution Properties

Notation Let k be an *alphabet size*, and denote by Δ_k the collection of distributions over alphabet $[k] := \{1, \ldots, k\}$. Let $d \in \mathbb{Z}^+$ be a *dimension parameter*. Let $p := (p_1, \ldots, p_d)$ be a d-tuple of

distributions in Δ_k, and for each $j \in [k]$, denote by $p(j) := (p_1(j), \ldots, p_d(j))$ the vector of the probabilities associated with j. A d-dimensional *multi-distribution property* over the alphabet $[k]$ is a mapping (functional) $f : \Delta_k^d \to \mathbb{R}$. Below, we present six examples arising in vital applications.

DISTRIBUTION MIXTURE TESTING

There has been an extensive study on testing properties of distributions over large domains in the past decades. For example, given sample access to an unknown distribution over $[k]$, a sequence of research works (Batu et al., 2000, 2001; Paninski, 2008; Goldreich and Ron, 2011; Chan et al., 2014; Acharya et al., 2015; Diakonikolas and Kane, 2016; Hao and Orlitsky, 2019b) address the problem of testing whether the distribution is uniform or ε far from it in ℓ_1 distance. For another example, one can relax the requirement of exactness in the previous example and ask whether the distribution is at most $\varepsilon/2$ or at least ε far from the uniform (Valiant and Valiant, 2011a,b; Jiao et al., 2018; Hao et al., 2018; Hao and Orlitsky, 2019a,b,c). While the former problem is known as "uniformity testing", the latter is often referred to as "tolerant uniformity testing" (Canonne, 2015).

Recently, the work of Aliakbarpour et al. (2019) takes a different perspective of the problem to consider a setting that involves three distributions, and asks if one can distinguish between the case where one distribution is the mixture of the other two, and that where it is ε-far from any such mixtures. In particular, they show that for several scenarios considered in the paper, a sample size that is sub-linear in k is sufficient. Following the previous discussions, we can consider a "tolerant" version of this problem. Specifically, we aim to estimate the quantity

$$M(p) := \min_{\alpha \in [0,1]} |p_1 - \alpha p_2 - (1-\alpha) p_3|$$

to the desired accuracy ε. This setting can accommodate more than three distributions. Concretely, keep the same notation and set p to be a d-distribution tuple in Δ_k^d. The property generalizes to

$$M(p) := \min_{\alpha \in \Delta_{d-1}} |p_1 - \sum_{i=1}^{d-1} \alpha(i) \cdot p_{i+1}|.$$

DISTRIBUTION DIVERGENCES

Distribution divergences quantify the similarity between related data sources in numerous learning applications, such as classification, testing, and Bayesian inference (Batu et al., 2000, 2001; Ron, 2010; Cover and Thomas, 2012; Blei et al., 2017). As these quantities reflect the fundamental limits of inference, it is worth obtaining accurate estimates of their values under various settings. Several recent works in property estimation (Valiant and Valiant, 2011a; Han et al., 2016; Acharya, 2018; Bu et al., 2018; Jiao et al., 2018; Charikar et al., 2019) consider designing *min-max estimators* that have the best worst-case guarantees for some f-divergences, including the ℓ_1 distance and KL-divergence, when both distributions are unknown. Utilizing the theory developed in this paper, we address another two essential divergences that are under-explored.

ℓ_q distance with $q \geq 1$ The first divergence measure we consider is the well-known ℓ_q distance. Note that in particular, this covers the ℓ_1 distance, the single instance also belonging to the class of f-divergences. Formally, the ℓ_q distance between any two distributions p_1 and p_2 is

$$\ell_q(p_1, p_2) := \sum_{j \in [k]} |p_1(j) - p_2(j)|^q.$$

Under the distributional setting, it is classical to employ (near-) unbiased estimators for integer q. Concurrently, there is a rich literature of work on approximating the ℓ_q distance, e.g., for $q \leq 2$ (Indyk, 2006; Li, 2007; Li and Hastie, 2007; Li, 2008) and for $q = 4, 6, 8, ...$ (Li et al., 2010).

Triangular discrimination The second divergence measure is *triangular discrimination* (Topsøe, 2000; Lu and Li, 2015), defined for two distributions p_1 and p_2 as

$$\Gamma(p_1, p_2) := \sum_{j\in[k]} \frac{(p_1(j) - p_2(j))^2}{p_1(j) + p_2(j)}.$$

The triangular discrimination is equivalent to the well-known harmonic mean divergence (T., 2006), and is analogous to the Chi-squared divergence yet does not become infinity at zero probabilities. Interestingly, in the literature, the triangular discrimination is often called the "Chi-squared distance" (Li et al., 2008; Wang et al., 2009; Vedaldi and Zisserman, 2012; Li et al., 2013).

INTERSECTION KERNEL AND SUMS OF MINS

As illustrated above, the ℓ_q divergence is a generalization of the commonly used ℓ_1 distance. Below, we consider a different generalization that applies to cases involving multiple distributions. To begin with, note that the ℓ_1 distance between two distributions p_1 and p_2 satisfies

$$\sum_{j\in[k]} |p_1(j) - p_2(j)| = \sum_{j\in[k]} (p_1(j) + p_2(j) - 2\min\{p_1(j), p_2(j)\}) = 2 - 2\sum_{j\in[k]} \min\{p_1(j), p_2(j)\}.$$

The last part on the right-hand side corresponds to the *intersection kernel*,

$$K_I(p_1, p_2) := \sum_{j\in[k]} \min\{p_1(j), p_2(j)\}.$$

which is popular in computer vision applications (Maji et al., 2008; Szeliski, 2010; Barla et al., 2003; Boughorbel et al., 2005) and its generalized version applies to generic classification tasks (Li, 2016). From the above equation, we see that K_I over Δ_k^2 measures the similarity between distributions and is equivalent to the ℓ_1 distance. Analogously, for any d-tuple of distributions $p \in \Delta_k^d$, we can define the intersection kernel of p as

$$K_I(p) := \sum_{j\in[k]} \min_{i\in[d]} p_i(j).$$

The value of $K_I(p)$ is at most 1, achieved iff p_i's are identical, and is at least 0, attained iff the distributions' support sets have no intersection.

COMMON SUPPORT COVERAGE OF MULTIPLE POPULATIONS

Given a sampling parameter m, the m-sample *support coverage* of any discrete distribution p_0 is

$$S_m(p_0) := \sum\nolimits_j \left(1 - (1 - p_0(j))^m\right),$$

the expected number of distinct symbols that will appear in a sample from p_0 of size m. Studied in Orlitsky et al. (2016); Acharya et al. (2017); Hao et al. (2018); Hao and Orlitsky (2019a,b,c), the task of estimating S_m is equivalent to the well-known unseen species problem – a classical task in ecology concerning the prediction of the number of species in an ecosystem not appeared in the observed sample (Good, 1953; Chao, 1984, 2004; Smith and van Belle, 1984). In this work, we introduce and study a natural generalization – *common support coverage*:

$$S_m(p) := \sum_j \prod_{i=1}^{d} \left(1 - (1 - p_i(j))^m\right),$$

the expected number of distinct symbols that appear at least once in all samples.

We motivate the study of this quantity with the following example. Imagine an ecologist who has access to butterfly samples from four islands that are geographically close to each other. A natural experiment is to compare the attributes of the same butterfly species on different islands, which intuitively shows how the habitats affect the species. To perform such a comparison, it is necessary to observe or capture at least one instance, a butterfly in this case, of the species from every island. Hence, with a size-m sample from each island, the expected number of such comparisons that can be made is exactly $S_m(p)$. In other words, the common support coverage reflects how many useful data points (tuples) the ecologist expects to have after a certain amount of work.

HIGH-DIMENSIONAL INDEPENDENCE TESTING

Besides testing mixtures of distributions, a frequently encountered inference task in data analysis is testing the independence of random variables (Batu et al., 2001; Alon et al., 2007; Rubinfeld and Xie, 2010; Levi et al., 2013; Canonne, 2015; Diakonikolas and Kane, 2016). The formulation we consider here again falls into the category of tolerance testing, and our aim is to distinguish between the case where the joint distribution is close to the product of marginals in the ℓ_1 distance, and the case where the distance is relatively large.

Unlike the properties mentioned above, in practice, the marginals are often over alphabets of different sizes. Hence, it is desired to design algorithms that accommodate such scenarios. Formally, given sample access to a (joint) distribution $\tilde{p}$ over $[\boldsymbol{k}] := [k_1] \times \ldots \times [k_d]$ and a d-tuple of (marginal) distributions $p \in \Delta_{\boldsymbol{k}} := \Delta_{k_1} \times \ldots \times \Delta_{k_d}$, we aim to estimate

$$\ell_1\left(\tilde{p}, p^{\times}\right) := \sum_{\boldsymbol{j} \in [\boldsymbol{k}]} \left|\tilde{p}(\boldsymbol{j}) - p^{\times}(\boldsymbol{j})\right| := \sum_{\boldsymbol{j} \in [\boldsymbol{k}]} \left|\tilde{p}(\boldsymbol{j}) - \prod_{i=1}^{d} p_i(\boldsymbol{j}_i)\right|,$$

the ℓ_1 distance between the joint and the product of the marginals, to a desired accuracy ε. Note that this property is not additive since each $p_i(j)$ appears $\prod_{i' \neq i} k_{i'}$ times on the right-hand side, while for an additive property the number of appearance should be one.

3. Main Results

Definitions Let k be an *alphabet size*, and denote by Δ_k the collection of distributions over $[k] := \{1, \ldots, k\}$. Let $d \in \mathbb{Z}^+$ be a *dimension parameter*. Let $p := (p_1, \ldots, p_d)$ be a d-tuple of distributions in Δ_k, and for each $j \in [k]$, denote by $p(j) := (p_1(j), \ldots, p_d(j))$ the vector of the corresponding probabilities. A d-dimensional *multi-distribution property* over the alphabet $[k]$ is a mapping (functional) $f : \Delta_k^d \to \mathbb{R}$. A property f is *additive* (additively separable) if there exists a function sequence $\{f_j : \mathbb{R}^d \to \mathbb{R}\}_{j=1}^{k}$ satisfying

$$f(p) = \sum\nolimits_j f_j(p(j)).$$

Let $n := (n_1, \ldots, n_d)$ be a sequence of *sampling parameters*. Draw a sample $X_i^{n_i} \sim p_i$ for each $i \in [d]$, and denote $X^n := (X_1^{n_1}, \ldots, X_d^{n_d})$, which we refer to as a *sample* from p and write

$X^n \sim p$. For each $j \in [k]$, denote by $N_{i,j}$ the number of times symbol j appearing in $X_i^{n_i}$. Correspondingly, we denote by $\hat{p}_i$ the empirical distribution that assigns each symbol $j \in [k]$ a probability $\hat{p}_i(j) := N_{i,j}/n_i$, and write $\hat{p} := (\hat{p}_1, \ldots, \hat{p}_d)$ as the *empirical distribution* of X^n. While p is unknown, we can infer the value of $f(p)$ by an *estimator* $\hat{f}$ mapping each sample from p to a real value. For example, the commonly used *empirical estimator* $\hat{f}^E$ estimates $f(p)$ by

$$\hat{f}^E(X^n) := f(\hat{p}).$$

As illustrated above, we aim to emulate the performance of the empirical estimator with an estimator that uses samples of smaller sizes. Equivalently, let $a := (a_1, \ldots, a_d)$ be an *amplification vector*, where $a_i > 1, \forall i \in [d]$. For $m := a * n = (a_1 n_1, \ldots, a_d n_d)$ and $Y^m \sim p$, we want to derive an estimator $\hat{f}$ such that $\hat{f}(X^n)$ is close to $\hat{f}^E(Y^m)$ under certain interpretations.

Assumptions The first problem we address is estimating an additive multi-distribution property f. Throughout the paper, we denote by $\mathscr{C}D$ the collection of continuous real functions over some domain D, and make the following two assumptions on $f = (f_j)_{j=1}^k$:

1. **[Lipschitzness]** $\forall j \in [k]$, function $f_j \in \mathscr{C}[0,1]^d$ is 1-Lipschitz regarding each of its inputs.
2. **[Regularity]** $\forall j \in [k]$, we have $f_j(p(j)) = 0$ if $p_i(j) = 0$ for any $i \in [d]$.

While being Lipschitz is, perhaps the simplest and most natural assumption one could make about function smoothness, the following lemma shows that we can express any multivariate real function as a sum of functions satisfying the regularity condition.

Lemma 1 *For any multivariate function $g : \mathbb{R}^d \to \mathbb{R}$,*

$$\tilde{g}\left((x_i)_{i=1}^d\right) := \sum_{j=0}^{d} \sum_{1 \le t_1 < \ldots < t_j \le d} (-1)^j \cdot g\left((x_i)_{i=1}^d|_{x_{t_s}=0, \forall s \in [j]}\right)$$

is a function satisfying $\tilde{g}\left((x_i)_{i=1}^d\right) = 0$ if $x_i = 0$ for any $i \in [d]$.

Theorems and corollaries Below, we present the major theorems and their corollaries. In terms of the coverage of statistical models, these results work for all additive multi-distribution properties that satisfy the regularity and Lipschitz conditions, and specific properties including those presented in the last section. In terms of the reach of methods, we study both the empirical estimator that achieves sample complexities linear in the alphabet size by our reasoning, and a new class of estimators (Section 4) that achieve sub-linear sample complexities. It is worth mentioning that for nearly all the problems considered here, sub-linear sample-complexity bounds are not known before our work, even for the three-distribution cases. Equally importantly, with the exception of Corollary 3, all proposed algorithms are near-linear-time computable in the sample sizes.

For the sake of clarity, we assume that the sample sizes n_i's are *equal* unless otherwise specified, and defer the generalization of these results to the case where n_i's are different to the appendices. In this equal-sample-size regime, we slightly abuse the notation and use n to denote both the vector $n := (n_1, \ldots, n_d)$ and each n_i. Whether n is a vector or a scalar will be clear from the context. Analogously, we assume that a_i's, the amplification factors, (resp. m_i's, the amplified sample sizes,) are equal and write a (resp. m) for both the vector and each a_i (resp. m_i). Note that $m = a \cdot n$ holds under both the vector and scalar interpretation.

In the following, we denote: the property of interest by f, which is additive and satisfies the Lipschitz and regularity conditions unless otherwise specified; our estimator by $\hat{f}$, which has an

explicit form (Section 4) and is near-linear-time computable; the empirical estimator of f by $\hat{f}^E$, which computes the sample empirical distribution and evaluates the property at this distribution.

Our first theorem and its corollary characterize the performance of our new estimator $\hat{f}$. Specifically, the theorem shows that our estimator amplifies the size of the data by a nontrivial factor of $\mathcal{O}(\sqrt{\log n})$ comparing to the empirical estimator, for every distribution tuple $p \in \Delta_k^d$. The corollary following it presents a weaker result under the classical sample-complexity formulation.

Theorem 1 *For any $a \geq 2.5$, $\tau \geq 1$, and $p \in \Delta_k^d$, if $\frac{2\log d}{\tau} \leq 6a \leq \sqrt{\frac{\log n}{\tau d}}$,*

$$\Pr_{X^n \sim p}\left(\left|\hat{f}(X^n) - \mathop{\mathbb{E}}_{Y^m \sim p}\left[\hat{f}^E(Y^m)\right]\right| \geq \frac{4d}{\sqrt{\tau}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{n^{1/6}}\right).$$

Note that τ can be any real value that satisfies the constraints. In particular, we can set τ to be an increasing function of n, then the estimation error $4d/\sqrt{\tau}$ will decrease as n increases. The proof of this theorem appears in Appendix J and follows by the results in earlier sections (see Section 4 for an outline). The next result is a corollary of Theorem 1 and Lemma 3 in Appendix E.

Corollary 1 *For any $\varepsilon > 0$, sufficiently large k, and $p \in \Delta_k^d$, if $n = \Omega\left(\frac{kd^{7/2}}{\sqrt{\log k}\varepsilon^3}\right)$,*

$$\Pr_{X^n \sim p}\left(\left|\hat{f}(X^n) - f(p)\right| \geq \varepsilon\right) = \tilde{\mathcal{O}}\left(\frac{1}{n^{1/6}}\right).$$

To demonstrate the power of the results, we claim that four of the properties presented in Section 2, ℓ_q distance, triangular discrimination, intersection kernel, and generalized support coverage, all satisfy the regularity and Lipschitz conditions after simple modifications. Hence, both Theorem 1 and Corollary 1 hold for these properties. See Appendix C for details and proofs.

Our second theorem characterizes the performance of the empirical estimators under only the Lipschitz assumption. In the large-alphabet regime, our result has the optimal dependence (up to constant factors) on all the parameters without any additional conditions.

Theorem 2 *Let f be an additive property satisfying the Lipschitz assumption. For any real $\varepsilon > 0$, $\delta \in (0, 1/2)$, and $p \in \Delta_k^d$, if $n \geq 4(kd^2 + d\log(1/\delta))/\varepsilon^2$,*

$$\Pr\left(\left|\hat{f}^E(X^n) - f(p)\right| \geq \varepsilon\right) \leq \delta.$$

The proof of this theorem appears in Appendix E. To see optimality, let u_k denote the uniform distribution over $[k]$, and consider the special case where $p = (u_k, \ldots, u_k)$ and $f(q) := \sum_i \ell_1(q_i, u_k) = \sum_i \sum_j |q_i(j) - 1/k|$ (Hence, the property value $f(p) = 0$, but the empirical estimator has no knowledge of this). By construction, the property f is additive and satisfies the Lipschitz assumption. Let $\hat{u}_k$ denote the empirical distribution of $Y^n \sim u_k$. Because of symmetry in our choice of p, the probability of $|\hat{f}^E(X^n) - f(p)| \leq \varepsilon$ is equal to that of $|\ell_1(\hat{u}_k, p)| \leq \varepsilon' := \varepsilon/d$. It is a standard result that the latter requires a sample of size $\Omega(k/\varepsilon'^2)$, establishing the desired optimality. This also shows that the empirical estimator cannot achieve the sub-linear sample complexity in Theorem 1.

High-dimensional independence testing As illustrated in the introduction, the ℓ_1 distance between the joint distribution and the product of marginals is not additive. Yet utilizing the structure of the property, we can still apply our learning approach to this fundamental task. The next two theorems characterize the performance of the empirical estimator and our estimator, showing that the problem again admits a sub-linear sample complexity bound.

More formally, given independent samples $Y^{\tilde{n}} \sim \tilde{p} \in \Delta_{[\boldsymbol{k}]}$ and $X^n \sim p \in \Delta_{\boldsymbol{k}}$, our aim is to estimate the quantity $\ell_1\left(\tilde{p}, p^{\times}\right)$. Observe that $\ell_1\left(\tilde{p}, p^{\times}\right)$ is 1-Lipschitz regarding every $\tilde{p}(\boldsymbol{j})$ and every $p_i(\boldsymbol{j}_i)$, for all $\boldsymbol{j} \in [\boldsymbol{k}]$. Denote by $\breve{p}$ and $\hat{p}$ the empirical distribution of $Y^{\tilde{n}}$ and X^n, respectively. Then the empirical estimator $\ell_1\left(\breve{p}, \hat{p}^{\times}\right)$ satisfies

Theorem 3 *Under the conditions presented above, for any $\varepsilon > 0$, $\delta \in (0, 1/2)$, $\tilde{n} \geq 8(2\prod_i k_i + \log(1/\delta))/\varepsilon^2$, and $n_i \geq 8(2k_i d^2 + d\log(1/\delta))/\varepsilon^2, \forall i \in [d]$,*

$$\Pr\left(\left|\ell_1\left(\breve{p}, \hat{p}^{\times}\right) - \ell_1\left(\tilde{p}, p^{\times}\right)\right| \geq \varepsilon\right) \leq \delta.$$

The proof of this theorem appears in Appendix E as well. The next theorem gauges the performance of our estimator, establishing a sub-linear sample complexity bound when the alphabet size k_i's do not differ from each other by too much.

Theorem 4 *Assume that $c_1 \log k_0 \leq \log k_i \leq c_2 \log k_0, \forall i \in [d]$, for some k_0 and absolute constants $c_1, c_2 > 0$. Then for any parameters $\varepsilon > 0$ and $d \in \mathbb{Z}^+$, sufficiently large k_0, and distributions $\tilde{p} \in \Delta_{[\boldsymbol{k}]}$ and $p \in \Delta_{\boldsymbol{k}}$, if $\tilde{n} = \Omega\left(\frac{(\prod_i k_i)d^{1/2}}{\sqrt{\log(\prod_i k_i)}\varepsilon^3}\right)$ and $n_i = \Omega\left(\frac{k_i d^{7/2}}{\sqrt{\log k_i}\varepsilon^3}\right), \forall i \in [d]$,*

$$\Pr_{Y^{\tilde{n}} \sim \tilde{p}, X^n \sim p}\left(\left|\hat{f}(Y^{\tilde{n}}, X^n) - \ell_1\left(\tilde{p}, p^{\times}\right)\right| \geq \varepsilon\right) = \tilde{\mathcal{O}}\left(\frac{1}{\tilde{n}^{1/6}}\right).$$

In fact, we establish a stronger result similar to Theorem 1 (see Theorem 5 in Appendix L), and we present one of its corollaries above for the ease of illustration. Furthermore, $\tilde{p}$ can be any distribution in $\Delta_{[\boldsymbol{k}]}$, and is unnecessary to have marginal p. We prove this theorem in Appendices K and L.

Tolerant mixture testing The next result shows that our approach yields the first sub-linear sample-complexity tester (or estimator) for tolerant distribution mixture testing. We first present a corollary of Theorem 1 involving three distributions, and then extend this to multi-distribution cases.

Corollary 2 *For any $\varepsilon > 0$, sufficiently large k, and $n = \Omega(k/(\varepsilon^3\sqrt{\log k}))$, given a sample X^n from p, we can compute an estimate $\hat{M}(X^n)$ in time $\tilde{\mathcal{O}}(n/\varepsilon)$ such that*

$$\left|\hat{M}(X^n) - M(p)\right| \leq \varepsilon,$$

with probability at least $9/10$. Simultaneously, the algorithm also provides an estimate of α that, when plugged into $|p_1 - \alpha p_2 - (1-\alpha)p_3|$, approximates $M(p)$ to an accuracy of ε.

In Appendix D, we provide a constructive proof of this corollary. For the multi-distribution case,

Corollary 3 *For any $\varepsilon > 0$, $d \in \mathbb{Z}^+$, sufficiently large k, and $p \in \Delta_k^d$, if $n = \Omega\big(kd^{7/2}/(\varepsilon^3\sqrt{\log k})\big)$, given a sample X^n from p, we can compute an estimate $\hat{M}(X^n)$ such that*

$$\left|\hat{M}(X^n) - M(p)\right| \leq \varepsilon,$$

with probability at least $9/10$ and in time $\tilde{\mathcal{O}}(dn \cdot (d/\varepsilon)^{d-1})$. The algorithm also provides an estimate $\alpha \in \Delta_{d-1}$ that, when plugged into $|p_1 - \sum_{i \in [d-1]} \alpha(i) \cdot p_{i+1}|$, approximates $M(p)$ to within ε.

4. Our Techniques and Estimators

Our first theoretical contribution, Theorem 1, is a nontrivial generalization of the main result in Hao et al. (2018), which addressed the special case of $d = 1$. The key component of the learning method in Hao et al. (2018) is an approximation technique based on the Bessel functions of the first kind. This technique approximates a single Poisson probability by a sum of such terms with smaller mean parameters, and achieves a small approximation error by re-weighting each term in the sum via Bessel functions, which are uniformly bounded and relatively smooth. In Appendix F, we present and analyze a variant of the estimator in Hao et al. (2018) which is not sensitive to sample changes. Henceforth, we refer to this technique as *Bessel smoothing*, since both its construction and analysis strongly rely on the Bessel functions and their attributes.

To extend the result to the multi-distribution setting, a natural idea is to apply the Bessel smoothing recursively to each component (i.e., a distribution in the d-tuple) of the property and utilize the resulting estimator. This is essentially the approach we adopt in the current work, yet as in many other theoretical analyses, showing the effectiveness of a natural approach is often nontrivial. Just as one may expect, the difficulty comes from both analyzing the bias and variance (or deviation probability) of the resulting estimator.

In the bias analysis, while the original function may take a simple form, once we apply the Bessel smoothing to a distribution component, the expectation of the resulting estimator becomes much more complicated. More importantly, for it to be reasonable to apply the smoothing technique to a consecutive component, the aforementioned expectation must be a property satisfying certain conditions required by the technique to work. Our first technical contribution is showing that the expectation induced by the Bessel smoothing is a linear operator over continuous functions that preserves regularity and *essentially* preserves Lipschitzness. Then we leverage this fact to bound the bias of our proposed estimator. See Appendices G and H for details.

As for the variance, the recursive application of Bessel smoothing yields a sophisticated estimator expression that mixes different products of statistics from multiple samples. While these products are distinct, they may share common factors, resulting in nontrivial dependence relations. To obtain tight variance upper bounds and handle the underlying dependency, we apply the law of total variance to separate the randomness associated with each distribution sample. This decomposition enables us to derive a recursion relation between the variances of the estimator's conditional expectation of different orders, where the word "order" refers to the amount of randomness that the conditioning is over. See Appendix I and specifically Appendix I.2 for details.

Our second theoretical contribution, Theorem 4, addresses the fundamental task of tolerant independence testing. As noted before, the property is non-additive, and hence, the previous analysis does not directly apply to this setting. A key observation is that the property is additive if we view it as a property of any of its input distribution components. Furthermore, for any single probability in the property's expression, the terms containing it are all Lipschitz functions with Lipschitz constants summing to one. Correspondingly, our second technical contribution shows that the linear operator induced by the Bessel smoothing not only preserves the Lipschitzness attribute of the input function, but also essentially preserves the magnitude of the Lipschitz constant (Appendix K).

Finally, we want to re-emphasize that sub-linear sample-complexity algorithms are not known in literature for tasks involving more than two distributions considered in the current work. In addition, except for tolerant mixture testing, our estimators are computable in time near-linear in the sample size, which is desired for large-domain applications.

We conclude this section by providing an explicit description of our estimator for d-distribution additive properties. In this case, the property of interest is

$$f(p) = \sum\nolimits_j f_j(p(j)).$$

Our strategy is to: 1) view the method in Hao et al. (2018) as a linear operator over $\mathscr{C}[0,1]$; 2) apply a variant of it (shown below) to each argument of (every) f_j while holding the other arguments fixed; 3) repeat this procedure until all probabilities are effectively replaced by their associated sample versions; 4) output the sum of our estimates for $f_j(p(j))$'s.

More concretely, for each $i \in [d]$, we randomly split the sample X_i^n from distribution p_i into two halves of equal size, and denote the empirical counts of each symbol j in the first and second halves by $N_{i,j}$ and $N'_{i,j}$, respectively. Then, for each $j \in [k]$ and each $i \in [d]$, we view f_j as a univariate function of its input $p_i(j)$ with all other parameters fixed, and apply the following operator with X and X' replaced by $N_{i,j}$ and $N'_{i,j}$, recursively. For parameters $n, a, m, \tau \geq 1$ in Theorem 1, the operator below maps a continuous function g to a function of random variables X and X':

$$\hat{H}(g, X, X') := \mathbb{1}_{X' \leq \tau} \cdot \left(\sum_{u=0}^{\tilde{u}} h^u_{X-u} \cdot g\left(\frac{u}{m}\right) \right) + (1 - \mathbb{1}_{X' \leq \tau} \mathbb{1}_{X \leq 2r\tau}) \cdot g\left(\frac{X}{n}\right),$$

where $\tilde{u} := 5a\tau$, $r := 6\tilde{u}$, and

$$h^u_s := (2a)^u (1-2a)^s \binom{s+u}{u} \Pr\left(\mathrm{Poi}(r) > s + 2u\right).$$

Here, τ is the parameter appearing in Theorem 1 that controls the estimation error. In the above estimator (operator), τ also serves as a threshold to separate small- and large-probability symbols.

The first and second components of the estimator respectively approximate the contributions from symbols of $\mathcal{O}(\tau/n)$ and $\Omega(\tau/n)$ probability. On a high level, the estimator: 1) approximates the performance of the m-sample empirical estimator of $g(\cdot)$; 2) is a variant of empirical estimator; 3) is a weighted sum of terms like $g(u/m)$. The parameter $\tilde{u}$ truncates the first sum as it is unlikely for a small-probability symbol to appear many times in the sample; parameter r determines how the summation terms will be attenuated, and serves as a smoothing parameter. Appendix F presents the detailed construction. For a concise two-page summary, see Section 6 of Hao et al. (2018).

5. Prior Results and Technique Comparisons

5.1. Property Estimation

There is a long line of research on estimating the properties of distributions from their samples, dating back several decades to the works of Good (1953); Efron and Thisted (1976). Because of the wide applications of property estimation in multiple disciplines, it has attracted significant attention from researchers working on information theory (Jiao et al., 2015; Orlitsky et al., 2016; Acharya et al., 2017; Jiao et al., 2018; Hao et al., 2018; Hao and Orlitsky, 2019a,b,c), theoretical computer science (Batu et al., 2005; Valiant and Valiant, 2011a, 2016, 2017; Charikar et al., 2019), and statistics (Paninski, 2003, 2004; Cai and Low, 2005, 2011; Wu and Yang, 2016, 2019). Below, we classify the results into two categories according to the parameter regimes.

The classical regime addresses the case where the sample size is much larger than the dimension of parameters. For a large class of statistical models, it is known (Van der Vaart, 2000) that the empirical estimator that utilizes the sample empirical distribution performs optimally. However,

modern learning applications frequently encounter problems with parameter dimensions comparable or even larger than the number of observations available. The desire to design estimators that outperform the empirical estimator in this data-sparse regime has driven the research in property estimation for the last two decades. For brevity, below we focus on results in the latter regime.

Before moving on to discuss the literature, we emphasize that both Theorem 1 and Corollary 1 cover a broad class of non-symmetric properties in a unified manner. While some common properties including those presented in Section 2 are symmetric (and additive), the class of non-symmetric properties is clearly much more general and practically important. In particular, it is easy to extend a symmetric additive property to a non-symmetric one. For example, one can associate different weights with the domain symbols and re-weighting the corresponding functions, for example, the expectation and variance of a random variable. A common motivation for considering these properties is to reflect the unequal levels of importance of different symbols. From a technical point of view, to estimate a symmetric property, it suffices to either estimate the function value of each symbol or the distribution probability multiset. On the other hand, for non-symmetric properties, the distribution multiset is generally insufficient for the purpose of estimation. As we illustrate below, this leads to the failure of several common estimation techniques, thus making our work distinct.

Next, we review some of the major techniques in literature besides Bessel smoothing and argue that none of them can easily adapt to our setting to yield similar results.

Plug-in with PML While the empirical distribution maximizes the probability of obtaining the labeled sample, the profile maximum likelihood (PML) finds a distribution maximizing the probability of observing the unlabeled sample, i.e., the multiset of empirical symbol counts Orlitsky et al. (2004). The PML plug-in estimator of a symmetric property simply evaluates the property value of the PML estimate. In the case of $d = 1$, Acharya et al. (2017) show that PML plug-in is sample-optimal for four additive symmetric properties. More recently, Hao and Orlitsky (2019b) extend PML's optimality to any symmetric additive properties that are appropriately Lipschitz, Charikar et al. (2019) propose an efficiently computable variant of PML, and Hao and Orlitsky (2020) further derive non-trivial results on estimating any symmetric properties.

As for the case of $d > 1$, the paper of Acharya (2018) proposes a generalization of the original single-distribution PML approach and shows that if there is an estimator achieving an ε error with probability at least $1 - \exp(-\Theta_d(\max_i n_i^d))$, over all d-tuples of distributions, then the generalized PML will achieve twice this error with high probability. However, as demonstrated by the aforementioned papers (Paninski, 2003; Valiant and Valiant, 2011a, 2017; Jiao et al., 2015; Wu and Yang, 2016; Orlitsky et al., 2016; Hao and Orlitsky, 2019a,c), showing the existence of a "nice" estimator that has sub-linear sample complexities and exponentially concentrates around its mean value, is highly nontrivial (Acharya et al., 2017; Hao and Orlitsky, 2019b, 2020). For the case of $d > 1$, existing results imply only the existence of such estimators for some specific two-distribution divergences, for example the KL-divergence (Acharya, 2018; Charikar et al., 2019). Hence establishing the efficiency of the PML approach in general for $d \geq 2$ remains open.

A different Bessel-type smoothing method A different smoothing method based on Bessel functions is proposed in Orlitsky et al. (2016); Acharya et al. (2017), and later adapted in Raghunathan et al. (2017) to estimate a particular multi-distribution property that generalizes the support coverage (different from the one mentioned in Section 2). However, as we explained below, the method in Orlitsky et al. (2016); Acharya et al. (2017); Raghunathan et al. (2017) does not seem to generalize to most of the additive properties considered in Hao et al. (2018) or this paper.

More specifically, we again think about the $d = 1$ case, i.e., estimating the single-distribution support coverage. Due to additiveness and symmetry, the problem is essentially approximating the function $1 - (1 - p_i)^m \approx 1 - e^{-mp_j}$, where the "$\approx$" operator changes the whole property value by at most 2. By the well-known series expansion $1 - e^{-y} = -\sum_t (-y)^t/t!$, a naive approach is to construct an unbiased estimator for each term p_j^t, which is possible under *Poisson sampling* with sample size being an independent Poisson random variable. However, both theoretical analysis and experimental evidences quickly show that the resulting property estimator has an unsatisfiable variance. To address this issue, Orlitsky et al. (2016) propose a smoothing method that weights each component $(-y)^t/t!$ by a real weight w_t. In particular, based on an integral expression of $1 - e^{-y}$ involving the first-order Bessel function, we can set $w_t = \Pr(\mathrm{Poi}(r) \geq t)$ for some parameter $r > 0$ and choose r properly to balance the estimator's variance and squared bias. This estimator achieves a sample complexity sublinear in m. Subsequently, Raghunathan et al. (2017) apply the same smoothing scheme to a generalization of support coverage involving multiple distributions.

By its construction, the method in Orlitsky et al. (2016) aims at approximating the exponential function e^{-y}. Yet as the derivations in Section F.1 demonstrate, we need to approximate a sequence of Poisson probabilities, i.e., functions in the form of $e^{-y}y^j/j!$. Following the reasoning in Section F, we can show that if one applies the approximation method in Orlitsky et al. (2016) to the e^{-y} component, no parameter r can yield the desired bias while maintaining a vanishing variance. Indeed, this is one of the motivations for the authors of the support-coverage-estimation paper to propose the novel Bessel smoothing method in Hao et al. (2018), which utilizes Bessel functions of different orders to address different functions $e^{-y}y^j/j!$ and consequently general single-distribution properties. Therefore, we naturally adopt the method in Hao et al. (2018) to tackle the problem of multi-distribution property estimation.

Another two interesting methods appearing in the literature are the **plug-in with linear programming** (Valiant and Valiant, 2011a, 2016, 2017; Han et al., 2018) and **min-max polynomial approximation** (Jiao et al., 2015; Wu and Yang, 2016; Hao and Orlitsky, 2019c). For space considerations, we postpone relevant discussions and comparisons to Appendix B.

5.2. Property Testing

While the property estimation framework aims to infer the $f(p)$ value, (distribution) property testing examines whether the underlying distribution(s) possess a certain attribute or not. Often, the latter reduces to a statistical test between two hypotheses, $f(p) \in A$ and $f(p) \in B$, with $A \cap B = \emptyset$.

There is a rich literature on this topic. Interested readers can refer to the survey by Canonne (2015) for a thorough review of prior works and open problems. Besides the numerous references in Section 2, Batu et al. (2005) considered multiplicative entropy estimation from a testing perspective. Hence, a future research direction is extending our results to this multiplicative estimation setting.

From a technical viewpoint, many distribution testing papers, such as Batu et al. (2000) and Diakonikolas and Kane (2016), build their algorithm based on estimating the ℓ_2 distance, or equivalently, counting "symbol collisions". The induced algorithms are simple and geared towards specific testing tasks. In particular, the ℓ_2 scheme puts much attention on relatively large probabilities and are not suitable for many other tasks such as the ℓ_1-tolerant testing of distribution closeness.

On the other hand, one may expect these algorithms to be more efficient as they focus on a specific problem. Hence, another research direction is determining the optimal sample complexity for any of the problems studied in Section 2 and 3, which are fundamental and practically important.

References

Jayadev Acharya. Profile maximum likelihood is optimal for estimating KL divergence. In *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1400–1404, Vail, CO, 2018.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3591–3599, Montreal, Canada, 2015.

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, 2016.

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)*, pages 11–21, Sydney, Australia, 2017.

Maryam Aliakbarpour, Ravi Kumar, and Ronitt Rubinfeld. Testing mixtures of discrete distributions. In *Conference on Learning Theory (COLT)*, pages 83–114, Phoenix, AZ, 2019.

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the 39th Annual ACM symposium on Theory of Computing (STOC)*, pages 496–505, 2007.

Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proceedings of the 2003 International Conference on Image Processing (ICIP)*, pages 513–516, Barcelona, Spain, 2003.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedints of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, Redondo Beach, CA, 2000.

Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, Las Vegas, NV, 2001.

Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Sabri Boughorbel, Jean-Philippe Tarel, and Nozha Boujemaa. Generalized histogram intersection kernel for image recognition. In *Proceedings of the 2005 International Conference on Image Processing (ICIP)*, pages 161–164, Genoa, Italy, 2005.

Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pages 771–782, Portland, OR, 2015.

Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V. Veeravalli. Estimation of KL divergence: Optimal minimax rate. *IEEE Trans. Inf. Theory*, 64(4):2648–2674, 2018.

Jorge Bustamante. *Bernstein operators and their properties*. Birkhuser Basel, 2017.

T Tony Cai and Mark G Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005.

T Tony Cai and Mark G Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

Clément L. Canonne. A survey on distribution testing – Your data is big. But is it blue? *Online Survey Draft*, 2015.

A. G. Carlton. On the bias of information estimates. *Psychological Bulletin*, 71(2):108–109, 1969.

Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, Portland, OR, 2014.

Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.

Anne Chao. Species estimation and applications. *Encyclopedia of statistical sciences*, 12, 2004.

Anne Chao and Chun-Huo Chiu. Species richness: Estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, pages 1–26, 2014.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 780–791, Phoenix, AZ, 2019.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, 14(3):462–467, 1968.

Fan R. Chung and Linyuan Lu. *Complex graphs and networks*, volume 107. American Mathematical Soc., 2017.

Robert K. Colwell, Anne Chao, Nicholas J. Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L. Chazdon, and John T. Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1): 3–21, 03 2012.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2nd edition, 2012.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694, New Brunswick, NJ, 2016.

Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

Wulfram Gerstner and Werner M. Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation - In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75. Springer, 2011.

Irving John Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953.

Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, seventh edition, 2007.

Peter J. Haas, Jeffrey F. Naughton, S. Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *Proceedings of 21th International Conference on Very Large Data Bases (VLDB)*, pages 311–322, Zurich, Switzerland, 1995.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of KL divergence between discrete distributions. In *Proceedings of the 2016 International Symposium on Information Theory and Its Applications (ISITA)*, pages 256–260, Monterey, CA, 2016.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance. In *Conference On Learning Theory (COLT)*, pages 3189–3221, Stockholm, Sweden, 2018.

Y. Hao and A. Orlitsky. Data amplification: Instance-optimal property estimation. In *arXiv preprint arXiv:1903.01432*, To appear at (ICML 2020), 2019a.

Y. Hao and A. Orlitsky. Profile entropy: A fundamental measure for the learnability and compressibility of discrete distributions. In *arXiv preprint arXiv:2002.11665*, 2020.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10989–11001, Vancouver, Canada, 2019b.

Yi Hao and Alon Orlitsky. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11104–11114, Vancouver, Canada, 2019c.

Yi Hao, Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8848–8857, Montréal, Canada, 2018.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

Iuliana Ionita-Laza, Christoph Lange, and Nan M. Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences (PNAS)*, 106(13):5008–5013, 2009.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory*, 61(5):2835–2885, 2015.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Trans. Inf. Theory*, 64(10):6672–6706, 2018.

Ian Kroes, Paul W Lepp, and David A Relman. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences (PNAS)*, 96(25):14547–14552, 1999.

Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9:295–347, 2013.

Ping Li. Very sparse stable random projections for dimension reduction in $l_\alpha(0 < \alpha <= 2)$ norm. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 440–449, San Jose, CA, 2007.

Ping Li. Computationally efficient estimators for dimension reductions using stable random projections. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 403–412, Pisa, Italy, 2008.

Ping Li. Generalized intersection kernel. Technical report, arXiv:1612.09283, 2016.

Ping Li and Trevor Hastie. A unified near-optimal estimator for dimension reduction in $l_\alpha(0 < \alpha <= 2)$ using stable random projections. In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912, Vancouver, Canada, 2007.

Ping Li and Cun-Hui Zhang. A new algorithm for compressed counting with applications in shannon entropy estimation in dynamic data. In *The 24th Annual Conference on Learning Theory (COLT)*, pages 477–496, Budapest, Hungary, 2011.

Ping Li, Kenneth Ward Church, and Trevor Hastie. One sketch for all: Theory and application of conditional random sampling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 953–960, Vancouver, Canada, 2008.

Ping Li, Michael W. Mahoney, and Yiyuan She. Approximating higher-order distances using random projections. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 312–321, Catalina Island, CA, 2010.

Ping Li, Gennady Samorodnitsky, and John Hopcroft. Sign Cauchy projections and Chi-square kernel. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2571–2579, Lake Tahoe, NV, 2013.

Guoxiang Lu and Bingqing Li. A class of new metrics based on triangular discrimination. *Information*, 6(3):361–374, 2015.

Zachary F Mainen and Terrence J Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995.

Subhransu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 2008.

Chang Xuan Mao and Bruce G Lindsay. Estimating the number of classes. *The Annals of Statistics*, pages 917–930, 2007.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Donald R. McNeil. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 68(341):92–96, 1973.

Alon Orlitsky, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 426–435. AUAI Press, 2004.

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

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inf. Theory*, 50(9):2200–2203, 2004.

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

Christopher J. Quinn, Negar Kiyavash, and Todd P. Coleman. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Trans. Signal Process.*, 61(12):3173–3182, 2013.

Aditi Raghunathan, Gregory Valiant, and James Zou. Estimating the unseen from multiple populations. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2855–2863, Sydney, Australia, 2017.

Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends® in Theoretical Computer Science*, 5(2):73–205, 2010.

Ronitt Rubinfeld and Ning Xie. Testing non-uniform k-wise independent distributions over product spaces. In *Automata, Languages and Programming, 37th International Colloquium (ICALP)*, pages 565–581, Bordeaux, France, 2010.

Eric P Smith and Gerald van Belle. Nonparametric estimation of species richness. *Biometrics*, pages 119–129, 1984.

Otto Szasz. Generalization of S. Bernsteins polynomials to the infinite interval. *J. Res. Nat. Bur. Standards*, 45(3):239–245, 1950.

Richard Szeliski. *Computer vision: Algorithms and applications*. Springer Science & Business Media, 2010.

Inder Jeet T. Bounds on triangular discrimination, harmonic mean and symmetric Chi-square divergences. *Journal of Concrete & Applicable Mathematics*, 4(1), 2006.

Ronald Thisted and Bradley Efron. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory*, 46(4):1602–1609, 2000.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 403–412, Palm Springs, CA, 2011a.

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 ACM Symposium on Theory of Computing (STOC)*, pages 685–694, San Jose, CA, 2011b.

Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 142–155, Cambridge, MA, 2016.

Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *J. ACM*, 64(6):37:1–37:41, 2017.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Rob R de Ruyter van Steveninck, Geoffrey D Lewen, Steven P Strong, Roland Koberle, and William Bialek. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808, 1997.

Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492, 2012.

Gang Wang, Derek Hoiem, and David A. Forsyth. Building text features for object image classification. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1367–1374, Miami, FL, 2009.

George Neville Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. 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*, 47(2):857–883, 2019.

Appendix A. Outline

The appendices are organized as follows. Appendix B completes the discussion in Section 5. Appendix C shows that simple variants of several properties presented in Section 2 are regular and Lipschitz. Appendix D proves Corollary 2 on tolerant mixture testing. In Appendix E, we analyze the performance of the most widely used empirical estimators, establishing Theorem 2 and 3.

In Appendix F, we present and analyze a variant of the estimator in Hao et al. (2018), which serves as a basic tool for subsequent constructions and reasoning. Appendix G treats the expectation of the estimator as a linear operator over continuous functions, and then proceeds to analyzing the estimator's analytical attributes, showing how they imply the desired result for $d = 2$. Through mathematical induction, Appendix H completes the bias analysis in the proof of Theorem 1. Appendix I continues the proof and analyzes the variance of the modified estimator for $d = 1$, and then extends the analysis to $d > 1$ via the aforementioned decomposition via the law of total variance. Note that all these results hold without assuming that the sample size n_i's are equal. In Appendix J, we specialize the result to the equal-sample-size case in Theorem 1 and fully establish the theorem.

We then proceed in Appendix K to the proof of Theorem 4 on the tolerant testing of high-dimensional independence. Despite that the property is not additive, we argue that the linear operator induced by the proposed estimator nearly preserves the Lipschitz constant of the original property with respect to each of its arguments. Following this claim, we analyze the bias of the estimator in Appendix K.1 and relate this bias to that of the empirical estimator having access to more observations. Utilizing the same claim and a tight bound on the estimator's coefficients, we establish in Section K.2 upper bounds on the estimator's sensitivity, i.e., the maximum difference in the estimator's values for two inputs differing at exactly one location. Finally, regardless of the involved statistical dependency, McDiarmid's inequality (Lemma 22) shows that the estimator is highly concentrated around its mean value. Consolidating these results yields Theorem 4.

Appendix B. Comparisons

This section continues our discussion in Section 5.

Plug-in with linear programming The linear-programming based methods, initiated by Efron and Thisted (1976), and analyzed and extended in the work of Valiant and Valiant (2011a, 2016, 2017) and its refinement Han et al. (2018), essentially estimates the moments of the underlying distribution from the samples, and through linear-programming, finds a distribution whose (low-order) moments are consistent with these estimates.

Three properties are considered in Valiant and Valiant (2016, 2017) and the corresponding estimators are shown to achieve optimal sample complexities for Shannon entropy, support size, and distance to uniformity in the constant error regime. The estimator proposed in Valiant and Valiant (2011a) uses similar techniques and achieves optimal sample complexity for Shannon entropy in terms of both the alphabet size and desired accuracy. We notice that Raghunathan et al. (2017) constructs a multi-distribution linear program and applies it to estimate a different generalization of the support coverage property. However, the paper evaluates only this linear program experimentally and provides no theoretical guarantees.

Applying such moment-based methods locally instead of globally, the work of Han et al. (2018) designs a refined estimator whose sample complexities are optimal for Shannon entropy, power sum, and support size, over broader error regimes. Nonetheless, it is not known if this method extends to

the multi-distribution case where $d > 2$. Consequently, it is also not clear what kind of guarantees this estimator will have on symmetric property estimation even if such an extension is possible. In addition, similar to most linear-programming based property estimators, the computation of this estimator takes polynomial time, thus may not be suitable for large-scale learning applications.

Min-max polynomial approximation For several single-distribution properties including entropy and support coverage, the empirical estimator performs well in estimating the function value of $f_j(p_j)$ unless p_j belongs to some sub-interval(s) of $[0, 1]$ where f_j is non-smooth, causing a non-negligible bias. The method of *min-max polynomial approximation* first estimates each probability p_j by its empirical frequency $\hat{p}_j$, and if $\hat{p}_j$ falls into the non-smooth segment of f_j, the method replaces the function by its local min-max polynomial approximation and finds an unbiased estimate for the polynomial, otherwise, it uses a simple bias-corrected variant of the empirical plug-in estimator. The method yields sample-optimal estimators for several properties involving one or two distributions, such as entropy, support size, ℓ_1 distance, KL-divergence (Jiao et al., 2015; Wu and Yang, 2016; Bu et al., 2018; Han et al., 2016; Jiao et al., 2018; Wu and Yang, 2019), and more generally, additive Lipschitz properties (Hao and Orlitsky, 2019c).

This is similar to our method in the sense that both methods start from the classical empirical plug-in estimator, and replace the inaccurate part of the empirical estimator with a polynomial-based estimator for bias correction. The major difference between the two methods, as one would expect, lies in the construction and analysis of the polynomial estimator.

For simplicity, we first think about $d = 1$. In Bessel smoothing, the approximation polynomial has coefficients being linear combinations of the function values at multiple points, where "function" refers to the univariate function to be approximate by polynomials. In the case of the min-max method, the polynomial is the min-max polynomial over certain interval that achieves the least maximum deviation from the function. Even for simple function classes, the mapping from a function to its min-max polynomial is not linear, and the coefficients do not admit closed-form formulas.

Now consider extending both methods to $d > 1$. For Bessel smoothing, by the above mentioned nice attributes established in the current paper, each step in the approximation process is essentially a linear operator that nearly preserves Lipschitzness of all continuous functions. Hence we can naturally apply this technique to each distribution component and obtain sub-linear sample estimators. On the other hand, for $d > 1$, it is well-known in approximation theory that the min-max polynomial is usually non-unique. In fact, even for ℓ_1 distance, a basic additive symmetric property involving only two distributions ($d = 2$), the paper (Jiao et al., 2018) argues that not all min-max polynomials will work. In particular, the paper also reasons that the min-max polynomial may yield only a sample complexity linear in the alphabet size. Consequently, the final construction in Jiao et al. (2018) utilizes the decomposition $|x - y| = |\sqrt{x} - \sqrt{y}| \cdot |\sqrt{x} + \sqrt{y}|$, approximates each of the two factors by its min-max polynomial, and employs the product of the resulting two polynomials. Given the involved construction, an extension to $d > 3$ and other properties seems to be quite nontrivial.

Appendix C. Regular and Lipschitz Properties

In the following, we prove the last claim made in Section 3, i.e., the ℓ_q distance, triangular discrimination, intersection kernel, and generalized support coverage, all satisfy the regularity and Lipschitz conditions after suitable modifications.

Distribution divergences We can decompose the ℓ_q distance $\ell_q(p_1, p_2)$ into three pieces such that every piece is regular and 1-Lipschitz regarding each of its arguments for $q \geq 1$. Specifically,

$$\ell_q(p_1, p_2) = \sum_{j\in[k]} \left(|p_1(j) - p_2(j)|^q - p_1(j)^q - p_2(j)^q\right) + \sum_{j\in[k]} p_1(j)^q + \sum_{j\in[k]} p_2(j)^q.$$

As for the triangular discrimination, we consider estimating the equivalent property

$$\frac{\Gamma(p_1, p_2) - 2}{4} = \sum_{j\in[k]} \left(\frac{(p_1(j) - p_2(j))^2}{4(p_1(j) + p_2(j))} - \frac{p_1(j) + p_2(j)}{4}\right).$$

It is sufficient to prove that $g(x, y) := (x-y)^2/(4x+4y) - (x+y)/4$ satisfies the two conditions. For any $y \in (0, 1]$, setting $x = 0$ implies $g(0, y) = y/4 - y/4 = 0$. In addition, the function's partial derivative with respect to x is $dg/dx = -y^2/(x+y)^2$, whose absolute value is at most 1 given $x, y \geq 0$. By symmetry, the function is 1-Lipschitz and regular regarding both arguments.

Intersection kernel Recall that the intersection kernel of a d-distribution tuple p is $K_I(p) := \sum_{j\in[k]} \min_{i\in[d]} p_i(j)$. The regularity of $K_I(p)$ follows by $\min_{i\in[d]} p_i(j) = 0$ iff $p_i(j) = 0$ for some $j \in [k]$. The Lipschitzness of $K_I(p)$ follows by $|\min\{x, y+z\} - \min\{x, y\}| = |\min\{0, (y-x)+z\} - \min\{0, (y-x)\}| \leq |z|$ where the inequality shows the Lipschitzness of the ReLU function.

Common support coverage While other properties considered above accept an upper bound of $\mathcal{O}(d)$, the d-distribution common support coverage can be arbitrarily large as m and k increase. Following the formulation in Orlitsky et al. (2016); Acharya et al. (2017), we normalize the property by its parameter m and consider

$$\tilde{S}_m(p) := \frac{S_m(p)}{m} = \sum_j \frac{1}{m} \prod_{i=1}^{d} \left(1 - (1 - p_i(j))^m\right).$$

We can verify that the normalized version is both Lipschitz and regular via simple algebra.

Appendix D. Tolerant Mixture Testing

In the following, we prove Corollary 2. To recap, we restate the corollary below.

Corollary 2 For any $\varepsilon > 0$, sufficiently large k, and $n = \Omega(k/(\varepsilon^3\sqrt{\log k}))$, given a sample X^n from p, we can compute an estimate $\hat{M}(X^n)$ in time $\tilde{\mathcal{O}}(n/\varepsilon)$ such that

$$\left|\hat{M}(X^n) - M(p)\right| \leq \varepsilon,$$

with probability at least $9/10$. Simultaneously, the algorithm also provides an estimate of α that, when plugged into $|p_1 - \alpha p_2 - (1-\alpha)p_3|$, approximates $M(p)$ to an accuracy of ε.

Proof For ease of exposition, we denote $M_\alpha(p) := |p_1 - \alpha p_2 - (1-\alpha)p_3|$ and $\alpha^* := \arg\min_\alpha M_\alpha(p)$. Decompose $M_\alpha(p)$ into several pieces (expressions in the square brackets below) so that all of them satisfy the regularity condition.

Then, we view α as a constant and apply our estimator to approximate each piece of

$$\begin{aligned}M_\alpha(p) =&|p_1 - \alpha p_2 - (1-\alpha)p_3| \\ =&[|p_1 - \alpha p_2 - (1-\alpha)p_3| - |\alpha p_2 + (1-\alpha)p_3| - |p_1 - \alpha p_2| \\ &-|p_1 - (1-\alpha)p_3| + p_1 + \alpha p_2 + (1-\alpha)p_3] \\ &+[|\alpha p_2 + (1-\alpha)p_3| - \alpha p_2 - (1-\alpha)p_3] + [|p_1 - \alpha p_2| - p_1 - \alpha p_2] \\ &+[|p_1 - (1-\alpha)p_3| - p_1 - (1-\alpha)p_3] + 2.\end{aligned}$$

Sum up the corresponding estimates and denote the sum by $\hat{M}_\alpha(X^n)$. The key observation is that for each multi-variate function component (i.e., f_j's) of the property, the value of our estimator is always a linear combination of the function evaluated at multiple points, and the coefficients for this linear combination and the evaluation points are completely determined by the input samples, amplification vector a, and parameter $\tau = \mathcal{O}(1/\varepsilon^2)$ (see Section 4). In addition, the linear combination has only a single nonzero coefficient. It is clear from the construction that computing this estimator only takes time near-linear in the input sample size.

Denote $J := \{j' \in \mathbb{N} : \varepsilon j' \in [0,1]\}$. Denote by $j^* \in J$ the index minimizing $|j\varepsilon - \alpha^*|$. Since $M_\alpha(p)$ is Lipschitz with respect to α, we obtain $|M_{\varepsilon j^*}(p) - M_{\alpha^*}(p)| \le \varepsilon$. By the triangle inequality, union bound, and Corollary 1, for $n = \Omega(k/(\varepsilon^3\sqrt{\log k}))$,

$$\min_{j\in J} \hat{M}_{\varepsilon j}(X^n) \le \hat{M}_{\varepsilon j^*}(X^n) \le M_{\varepsilon j^*}(p) + \varepsilon = (M_{\varepsilon j^*}(p) - M_{\alpha^*}(p)) + M_{\alpha^*}(p) + \varepsilon \le M_{\alpha^*}(p) + 2\varepsilon$$

and

$$\min_{j\in J} \hat{M}_{\varepsilon j}(X^n) \ge \min_{j\in J} M_{\varepsilon j}(p) - \varepsilon \ge M_{\alpha^*}(p) - \varepsilon,$$

with probability at least $1 - \tilde{\mathcal{O}}\left(1/(\varepsilon^{1/2}k^{1/6})\right)$. Let $\hat{j}$ denote the index that minimizes $\hat{M}_j(X^n)$. We output $\hat{\alpha} := \hat{j} \cdot \varepsilon$ as our estimate for α. Note that $\hat{\alpha}$ may not be close to α^*, yet the corresponding values of $M_\alpha(p)$ are close, which suffices for our purpose. ■

Appendix E. Performance of the Empirical Estimator

The *degree-t Bernstein polynomial* of a real function $g : [0,1] \to \mathbb{R}$ is defined as

$$B_t(g,x) := \sum_{j=0}^{t} \binom{t}{j} x^j (1-x)^{t-j} g\left(\frac{j}{t}\right).$$

Bernstein polynomials are closely related to empirical estimators. More specifically, draw a sample of size t from a Bernoulli random variable with success probability x. Denote by X the number of times symbol 1 appearing in the sample. The expected value of the empirical estimator for $g(x)$ is the Bernstein polynomial $B_t(g,x)$, i.e.,

$$B_t(g,x) = \mathbb{E}\left[g\left(\frac{X}{t}\right)\right].$$

The next lemma (Bustamante, 2017) (Proposition 4.9) shows that for a Lipschitz function, its Bernstein polynomial approximates the function well over $(0,1)$ and coincides with it at the boundary.

Lemma 2 *For any $g : [0,1] \to \mathbb{R}$ that is 1-Lipschitz, $t \geq 1$ and $x \in [0,1]$,*

$$|B_t(g,x) - g(x)| \leq \sqrt{\frac{x(1-x)}{t}}.$$

Utilizing this lemma, we bound the bias of the empirical estimator for an additive property.

Lemma 3 *Let f be an additive property satisfying the Lipschitz condition. For any distribution tuple $p \in \Delta_k^d$ and sampling vector n,*

$$\left| f(p) - \mathop{\mathbb{E}}_{X^n \sim p}[\hat{f}^E(X^n)] \right| \leq \sum_{i=1}^{d} \sqrt{\frac{k}{n_i}}.$$

Proof Let $\hat{p}$ be the empirical distribution associated with X^n. By the triangle inequality, we can decompose the absolute mean deviation of $\hat{f}^E$ into d parts,

$$\begin{aligned}
\left| f(p) - \mathbb{E}[\hat{f}^E(\hat{p})] \right| &\overset{(a)}{\leq} \sum_{i=1}^{d} |\mathbb{E}[f(p_1, \ldots, p_{d-i+1}, \hat{p}_{d-i+2}, \ldots, \hat{p}_d) - f(p_1, \ldots, p_{d-i}, \hat{p}_{d-i+1}, \ldots, \hat{p}_d)]| \\
&\overset{(b)}{\leq} \sum_{i=1}^{d} \sum_{j=1}^{k} |\mathbb{E}[f(p_1(j), \ldots, p_{d-i+1}(j), \hat{p}_{d-i+2}(j), \ldots, \hat{p}_d(j)) \\
&\qquad - f(p_1(j), \ldots, p_{d-i}(j), \hat{p}_{d-i+1}(j), \ldots, \hat{p}_d(j))]| \\
&\overset{(c)}{\leq} \sum_{i=1}^{d} \sum_{j=1}^{k} \sqrt{\frac{p_{d-i+1}(j)}{n_{d-i+1}}} \\
&\overset{(d)}{\leq} \sum_{i=1}^{d} \sqrt{\frac{k}{n_{d-i+1}}} \\
&\overset{(e)}{=} \sum_{i=1}^{d} \sqrt{\frac{k}{n_i}},
\end{aligned}$$

where both (a) and (b) follow by the triangle inequality; (c) follows by Lemma 2; (d) follows by the Cauchy-Schwarz inequality; (e) follows by a change of indices. ■

The next lemma, whose proof follows from McDiarmid's inequality (Lemma 22), demonstrates the concentration of empirical estimators around their mean values.

Lemma 4 *Let f be an additive property satisfying the Lipschitz condition. For any error parameter $\varepsilon > 0$ and distribution tuple $p \in \Delta_k^d$,*

$$\Pr_{X^n \sim p}\left(\left| \hat{f}^E(X^n) - \mathbb{E}[\hat{f}^E(X^n)] \right| \geq \varepsilon \right) \leq 2 \exp\left(-\frac{2\varepsilon^2}{\sum_i \frac{1}{n_i}} \right).$$

Proof By the Lipschitz condition, for any $i \in [d]$ and $j \in [n_i]$, changing a single observation $X_{i,j}$ changes the value of function g_i by at most $c_{i,j} := 1/n_i$. Then we have

$$\sum_i \sum_j c_{i,j}^2 = \sum_i n_i \cdot \frac{1}{n_i^2} = \sum_i \frac{1}{n_i}.$$

Hence by McDiarmid's inequality (Lemma 22),

$$\Pr_{X^n\sim p}\left(\left|\hat{f}^E(X^n)-\mathbb{E}[\hat{f}^E(X^n)]\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_i\frac{1}{n_i}}\right).$$

■

Utilizing the above lemma and setting $\varepsilon=\sqrt{\log(1/\delta)\sum_i(1/n_i)}$ for some $\delta\in(0,1/2)$, we characterize the performance of the empirical property estimator as follows.

Lemma 5 *Let f be an additive property satisfying the Lipschitz assumption. For any parameter $\delta\in(0,1/2)$, distribution tuple $p\in\Delta_k^d$, and sample $X^n\sim p$, with probability at least $1-\delta$,*

$$\left|\hat{f}^E(X^n)-f(p)\right|\leq\sum_{i=1}^d\sqrt{\frac{k}{n_i}}+\sqrt{\log\left(\frac{1}{\delta}\right)\sum_i\frac{1}{n_i}}.$$

When n_i's are equal, the above lemma implies

Lemma 6 *Under the conditions stated in Lemma 5, if $n\geq 4(kd^2+d\log(1/\delta))/\varepsilon^2$,*

$$\Pr_{X^n\sim p}\left(\left|\hat{f}^E(X^n)-f(p)\right|\geq\varepsilon\right)\leq\delta.$$

Testing high-dimensional independence Given independent samples $Y^{\tilde{n}}\sim\tilde{p}\in\Delta_{[\boldsymbol{k}]}$, and $X^n\sim p\in\Delta_{\boldsymbol{k}}$, the objective of tolerant independence testing is to estimate the quantity

$$\ell_1\left(\tilde{p},p^\times\right)=\sum_{\boldsymbol{j}\in[\boldsymbol{k}]}\left|\tilde{p}(\boldsymbol{j})-p^\times(\boldsymbol{j})\right|=\sum_{\boldsymbol{j}\in[\boldsymbol{k}]}\left|\tilde{p}(\boldsymbol{j})-\prod_{i=1}^d p_i(\boldsymbol{j}_i)\right|.$$

The key observation (Section K) is that $\ell_1\left(\tilde{p},p^\times\right)$ is 1-Lipschitz with respect to every $\tilde{p}(\boldsymbol{j})$ and every $p_i(\boldsymbol{j}_i)$, where $\boldsymbol{j}\in[\boldsymbol{k}]$. Denote by $\breve{p}$ and $\hat{p}$ the empirical distribution of $Y^{\tilde{n}}$ and X^n, respectively. Following the same derivation as in the proof of Lemma 4, for any $\varepsilon>0$,

$$\Pr\left(\left|\ell_1\left(\breve{p},\hat{p}^\times\right)-\mathbb{E}[\ell_1\left(\breve{p},\hat{p}^\times\right)]\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{\frac{1}{\tilde{n}}+\sum_i\frac{1}{n_i}}\right).$$

Correspondingly, the following result is an analogy to Lemma 3.

$$\left|\ell_1\left(\tilde{p},p^\times\right)-\mathbb{E}[\ell_1\left(\breve{p},\hat{p}^\times\right)]\right|\leq\sqrt{\frac{\prod_i k_i}{\tilde{n}}}+\sum_{i=1}^d\sqrt{\frac{k_i}{n_i}}.$$

Consolidating these results shows that the empirical estimator satisfies

Lemma 7 *Under the conditions stated above, for any $\varepsilon>0$ and $\delta\in(0,1/2)$, if $\tilde{n}\geq 8(2\prod_i k_i+\log(1/\delta))/\varepsilon^2$ and $n_i\geq 8(2k_id^2+d\log(1/\delta))/\varepsilon^2,\forall i\in[d]$,*

$$\Pr\left(\left|\ell_1\left(\breve{p},\hat{p}^\times\right)-\ell_1\left(\tilde{p},p^\times\right)\right|\geq\varepsilon\right)\leq\delta.$$

Poisson sampling For the sake of simplicity, we adopt the conventional *Poisson sampling* technique. That is, we make each of the sample sizes involved in the problem an independent Poisson random variable with mean value being equal to the targeting sample size. This does not change the nature of the problem or that of the estimators, and eliminates the dependence among symbols' empirical counts in the sample. More rigorously, in Appendix N, we show that for a d-distribution property and sampling vector $n = (n_1, \ldots, n_d)$, the expected values of the empirical estimator differ by only at most $3\sum_i n_i^{-1/3}$ under the two sampling models, i.e., fixed sampling and Poisson sampling, as long as all the sampling parameters are at least 44. As for the proposed estimator, we also assume that samples of independent Poisson sizes are provided. Below we argue that the same estimator will also work well in the fixed-sample-size case.

For clarity, we will further assume that all the n_i's are equal and suppress the sub-script i. According to the reasoning in Section K.2 and L (Theorem 5), our estimator for the property of high-dimensional independence is highly concentrated around the expected value of the larger-sample-size empirical estimator. The same reasoning works for our proposed estimator for general additive properties as well, showing that for sufficiently large n, the error-probability bound in Theorem 1 can be strengthened (with respect to n) to $d\exp(-n^{0.2})$. We would like to point out that this probability bound is generally not sufficient to establish the optimality of the aforementioned PML estimator (we need something like $\exp(-n^d)$), yet it is sufficient for establishing the efficiency of our estimator under the fixed sampling model. By simple algebra, the probability that a Poisson random variable with mean n will be exactly n is at least $1/(3\sqrt{n})$. Hence the probability that all the given d independent samples will have sample size n is at least $1/(3\sqrt{n})^d$. Therefore given that this event happens, the probability that the proposed estimator will violates the guarantee stated in Theorem 1 is at most $(3\sqrt{n})^d \cdot d\exp(-n^{0.2})$, which vanishes with n as long as $d \ll n^{0.2}/\log n$. As illustrated previously, d is often small for many applications. In addition, Theorem 1 assumes that $d \leq \log n$. Hence $d \ll n^{0.2}/\log n$ is not a strong assumption, implying the desired result.

Appendix F. Basic Case: $d = 1$

Let us first consider perhaps the simplest setting: Given a size-$\mathrm{Poi}(n)$ sample ($2n$ is chosen for the simplicity of notation) from a Bernoulli distribution with success probability x, our objective is to estimate the value of some real function f at point x and emulate the performance of an empirical plug-in estimator that has access to more sample points, i.e., a sample of size $m := na$ for some $a \geq 2.5$. Note that in this section, most of the parameters, e.g., m, n, and x, are 1-dimensional, i.e., a real number or an integer.

The basic design of our estimator follows from the construction in Hao et al. (2018), yet for the later analysis on the mean deviation probability, we need to modify the *large-probability estimator* for symbols with relatively high empirical frequencies. Besides this, we also tightened several deviation bounds, obtained non-asymptotic guarantees with fairly small constants, and simplified multiple proofs utilizing the theory of linear operators.

Formally, consider a real function $f \in \mathscr{C}[0,1]$ that is both 1-Lipschitz and regular. *We naturally extend the function and maintain its Lipschitzness by setting* $f(z) = f(1), \forall z \geq 1$.

Let x be an unknown real parameter in $[0,1]$. Given parameters $n \in \mathbb{N}$ and $a > 1$ satisfying $m := an \in \mathbb{N}$, and independent samples $X, X' \sim \mathrm{Poi}(\lambda)$ where $\lambda := nx$, we want to estimate

$$\boldsymbol{S}_m[f, x] := \mathop{\mathbb{E}}_{Y \sim \mathrm{Poi}(mx)} \left[f\left(\frac{Y}{m}\right) \right],$$

where $\boldsymbol{S}_m$ is the m-th order *Szász-Mirakyan operator* (See Section F.2). For a threshold $\tau > 0$ to be specified later, we can decompose this quantity into two parts:

$$\boldsymbol{S}_m[f,x] = \mathop{\mathbb{E}}_{X'\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X'\le\tau}] \mathop{\mathbb{E}}_{Y\sim\mathrm{Poi}(mx)}\left[f\left(\frac{Y}{m}\right)\right] + \mathop{\mathbb{E}}_{X'\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X'>\tau}] \mathop{\mathbb{E}}_{Y\sim\mathrm{Poi}(mx)}\left[f\left(\frac{Y}{m}\right)\right].$$

Due to the concentration of Poisson random variables, for a smooth function f and values of x that are small, the main contribution to $\boldsymbol{S}_m[f,x]$ comes from the first term on the right-hand side. Analogously, for relatively large x, the value of $\boldsymbol{S}_m[f,x]$ mainly depends on the second term. Consequently, we refer to the first term as the *small-probability part* $P_S(f,\lambda)$, and the second term as the *large-probability part* $P_L(f,\lambda)$.

First, we approximate the small-probability part $P_S(f,\lambda)$.

F.1. Small-Probability Estimator

Our estimator for P_S closely relates to the *Bessel functions of the first kind* (Gradshteyn and Ryzhik, 2007), which are solutions $J_v(y)$ of the *Bessel differential equation*. For notational convenience, further denote $f_u(y) := J_{2u}(2\sqrt{y})$.

Fixing $a > 1$, we define two functions. The first function takes $u \in \mathbb{Z}^+$ and $\lambda \in \mathbb{R}_{\ge 0}$ as input, and represents the probability of observing u when we sample from $\mathrm{Poi}(a\lambda)$:

$$h_a(u,\lambda) := \Pr(\mathrm{Poi}(a\lambda) = u) = \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \int_0^\infty e^{-\beta}\beta^u f_u(\beta(a-1)\lambda)d\beta,$$

where the last equality follows by Lemma 14 in the supplementary material of Hao et al. (2018). Truncating the inner integral at a level $r \in \mathbb{R}^+$ yields the second function

$$h_a^r(u,\lambda) := \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \int_0^r e^{-\beta}\beta^u f_u(\beta(a-1)\lambda)d\beta.$$

Note that $h_a(u,\lambda) = h_a^\infty(u,\lambda)$. For sufficiently large r, we naturally expect the function values of h_a and h_a^r to be close. The next lemma formalizes this intuition.

Lemma 8 *For any $\lambda \ge 0$, $a \ge 2.5$, and $r \ge 5(u+1) \vee 10(a-1)$,*

$$|h_a(u,\lambda) - h_a^r(u,\lambda)| \le e^{-\lambda}\lambda \cdot e^{-r/3}$$

and $|h_a^r(u,\lambda)| \le \Pr(\mathrm{Poi}(a\lambda) = u) + e^{-\lambda}\lambda \cdot e^{-r/3}$.

Proof By the inequality $J_{2u}(2\sqrt{y}) \le y/(u+1), \forall u \ge 1, y \ge 0$ (Watson, 1995; Hao et al., 2018), and the series expansion of the upper incomplete Gamma function, we have

$$\begin{aligned}
|h_a(u,\lambda) - h_a^r(u,\lambda)| &\le \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \int_r^\infty e^{-\beta}\beta^u |f_u(\beta(a-1)\lambda)|d\beta \\
&\le \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \int_r^\infty e^{-\beta}\beta^u \frac{\beta(a-1)\lambda}{u+1}d\beta \\
&\le e^{-\lambda}\lambda(a-1)\left(\frac{a}{a-1}\right)^u \frac{1}{(u+1)!}\int_r^\infty e^{-\beta}\beta^{u+1}d\beta \\
&= e^{-\lambda}\lambda(a-1)\left(\frac{a}{a-1}\right)^u \Pr(\mathrm{Poi}(r) \le u+1).
\end{aligned}$$

Further, by the Chernoff bound, for $a \geq 2.5$ and $r \geq 5(u+1) \vee 10(a-1)$,

$$e^{-\lambda}\lambda(a-1)\left(\frac{a}{a-1}\right)^{u-1}\Pr(\mathrm{Poi}(r) \leq u+1) \leq e^{-\lambda}\lambda \cdot \frac{r}{10} \cdot e^{(0.52r)/5} \cdot e^{-0.478r} \leq e^{-\lambda}\lambda \cdot e^{-r/3}.$$

Combined with the triangle inequality, the above derivation also yields

$$|h_a^r(u,\lambda)| \leq h_a(u,\lambda) + |h_a(u,\lambda) - h_a^r(u,\lambda)| \leq \Pr(\mathrm{Poi}(a\lambda) = u) + e^{-r/3}.$$

■

Now we consider estimating the small-probability part. By definition, we can rewrite P_S as

$$\Pr(\mathrm{Poi}(\lambda) \leq \tau) \mathop{\mathbb{E}}_{Y \sim \mathrm{Poi}(a\lambda)}\left[f\left(\frac{Y}{m}\right)\right] = \sum_{u=0}^{\infty}(\Pr(\mathrm{Poi}(\lambda) \leq \tau) \cdot \Pr(\mathrm{Poi}(a\lambda) = u)) \cdot f\left(\frac{u}{m}\right).$$

A naive unbiased estimator of this quantity suffers from large variance. For this reason, we first reduce the size of the sum to make it more manageable. Note that Poisson random variables highly concentrate around their mean values. Hence for parameter $u \gg a\tau$, the product $\Pr(\mathrm{Poi}(\lambda) \leq \tau) \cdot \Pr(\mathrm{Poi}(a\lambda) = u)$ becomes negligible. More concretely, the following lemma holds.

Lemma 9 *For any $\lambda, \tau \geq 0$, $a \geq 2.5$, and $u \geq 2.5a\tau$,*

$$\Pr(\mathrm{Poi}(a\lambda) \geq u) \cdot \Pr(\mathrm{Poi}(\lambda) \leq \tau) \leq \exp\left(-\frac{3}{8}\tau\right).$$

We postpone the proof of this lemma to Appendix M. Hence for $\tilde{u} \geq 2.5a\tau$, the magnitude of the partial sum over $u > \tilde{u}$ is at most

$$\begin{aligned}
&\Pr(\mathrm{Poi}(\lambda) \leq \tau)\left|\sum_{u=\tilde{u}+1}^{\infty}\Pr(\mathrm{Poi}(a\lambda) = u) f\left(\frac{u}{m}\right)\right| \\
\leq &\Pr(\mathrm{Poi}(\lambda) \leq \tau)\sum_{u=\tilde{u}+1}^{\infty} e^{-a\lambda}\frac{(a\lambda)^u}{u!}\frac{u}{m} \\
\leq &\frac{a\lambda}{m}\Pr(\mathrm{Poi}(\lambda) \leq \tau)\sum_{u=\tilde{u}+1}^{\infty} e^{-a\lambda}\frac{(a\lambda)^{u-1}}{(u-1)!} \\
= &\frac{\lambda}{n}\Pr(\mathrm{Poi}(\lambda) \leq \tau) \cdot \Pr(\mathrm{Poi}(a\lambda) \geq \tilde{u}) \\
\leq &\exp\left(-\frac{3}{8}\tau\right)\frac{\lambda}{n},
\end{aligned}$$

showing that we need to consider approximating only the first $\tilde{u}$ terms of the sum.

The last ingredient required for balancing the variance and bias is substituting each probability term $\Pr(\mathrm{Poi}(a\lambda) = u) = h_a(u,\lambda)$ by $h_a^r(u,\lambda)$. That is, we approximate the small-probability part by

$$H_S(f,\lambda) := \Pr(\mathrm{Poi}(\lambda) \leq \tau)\sum_{u=0}^{\tilde{u}} h_a^r(u,\lambda) \cdot f\left(\frac{u}{m}\right).$$

Our *small-probability estimator* is simply the unbiased estimator for this quantity. It will be clear later that reducing the truncation parameter r will generally increase the bias but decrease the variance of this estimator. Below we find the explicit form of our estimator for $P_S(f,\lambda)$.

Expanding the inner integral, we can rewrite $h_a^r(u,\lambda)$ as

$$\begin{aligned}
h_a^r(u,\lambda) &= \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \int_0^r e^{-\beta}\beta^u f_u(\beta(a-1)\lambda)d\beta \\
&= \frac{e^{-\lambda}}{u!}\left(\frac{a}{a-1}\right)^u \sum_{s=0}^{\infty}(-1)^s\frac{(\lambda(a-1))^{s+u}}{s!}\left(1-e^{-r}\sum_{j=0}^{s+2u}\frac{r^j}{j!}\right) \\
&= \sum_{s=0}^{\infty}\left(a^u(1-a)^s\binom{s+u}{u}\left(1-e^{-r}\sum_{j=0}^{s+2u}\frac{r^j}{j!}\right)\right)\left(e^{-\lambda}\frac{\lambda^{s+u}}{(s+u)!}\right) \\
&= \sum_{s=0}^{\infty}\left(a^u(1-a)^s\binom{s+u}{u}\Pr\left(\mathrm{Poi}(r)>s+2u\right)\right)\left(e^{-\lambda}\frac{\lambda^{s+u}}{(s+u)!}\right).
\end{aligned}$$

For brevity, we define the inner coefficients by

$$h_s^u := a^u(1-a)^s\binom{s+u}{u}\Pr\left(\mathrm{Poi}(r)>s+2u\right), \forall s\geq 0,$$

and set $h_s^u := 0, \forall s<0$. Then given $X\sim\mathrm{Poi}(\lambda)$, an unbiased estimator for $h_a^r(u,\lambda)$ is

$$\sum_{s\geq 0} h_s^u\cdot \mathbb{1}_{X=s+u} = h_{X-u}^u.$$

Consequently, our *unbiased* estimator for $H_S(f,\lambda)$, the small-probability estimator, is

$$\hat{H}_S(f,X,X') := \mathbb{1}_{X'\leq\tau}\cdot\left(\sum_{u=0}^{\tilde{u}} h_{X-u}^u\cdot f\left(\frac{u}{m}\right)\right).$$

Bounding the estimation bias Consolidating the previous results and applying Lemma 8, for $a\geq 2.5$, $\tau\geq 1$, $\tilde{u}\geq 2.5a\tau$, and $r\geq 5(\tilde{u}+1)\vee 10(a-1)$, the bias of $\hat{H}_S(f,X,X')$ is equal to

$$\begin{aligned}
|P_S(f,\lambda)-H_S(f,\lambda)| &\leq \Pr(\mathrm{Poi}(\lambda)\leq\tau)\left|\sum_{u=\tilde{u}+1}^{\infty}\Pr(\mathrm{Poi}(a\lambda)=u)f\left(\frac{u}{m}\right)\right| \\
&\quad + \sum_{u=0}^{\tilde{u}}|h_a^r(u,\lambda)-h_a(u,\lambda)|\cdot\left|f\left(\frac{u}{m}\right)\right| \\
&\leq \exp\left(-\frac{3}{8}\tau\right)\frac{\lambda}{n}+\sum_{u=0}^{\tilde{u}}\lambda\cdot\frac{u}{m} \\
&= \left(\exp\left(-\frac{3}{8}\tau\right)+\frac{\tilde{u}(\tilde{u}+1)}{2ae^{r/3}}\right)\frac{\lambda}{n} \\
&\leq \exp\left(-\frac{\tau}{3}\right)\frac{\lambda}{n}.
\end{aligned}$$

F.2. Large-Probability Estimator

In this section, we consider estimating the large-probability part:

$$P_L(f,\lambda) = \mathop{\mathbb{E}}_{X'\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X'>\tau}] \mathop{\mathbb{E}}_{Y\sim\mathrm{Poi}(a\lambda)}\left[f\left(\frac{Y}{an}\right)\right].$$

We approximate this quantity using the estimator $\hat{H}_L$ below. The reason for such construction and the choice of corresponding parameters will become clear in later sections, e.g., Section K.2. For now, we will take this estimator for granted and focus on analyzing its approximation attributes.

$$\hat{H}_L(f,X,X') := (1-\mathbb{1}_{X'\le\tau}\mathbb{1}_{X\le 2r\tau})\cdot f\left(\frac{X}{n}\right).$$

Analogous to the analysis in the last section, denote by $H_L(f,\lambda)$ the expectation of $\hat{H}_L(f,X,X')$. Then by the triangle inequality, the bias of this *large-probability estimator* is equal to

$$\begin{aligned}|P_L(f,\lambda)-H_L(f,\lambda)| \le &\left|\mathop{\mathbb{E}}_{X'\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X'>\tau}]\left(\mathop{\mathbb{E}}_{Y\sim\mathrm{Poi}(a\lambda)}\left[f\left(\frac{Y}{an}\right)\right]-\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(\lambda)}\left[f\left(\frac{X}{n}\right)\right]\right)\right| \\ &+\mathop{\mathbb{E}}_{X,X'\sim\mathrm{Poi}(\lambda)}\left|\mathbb{1}_{X'\le\tau}\mathbb{1}_{X>2r\tau}\cdot f\left(\frac{X}{n}\right)\right|.\end{aligned}$$

Noting that $\lambda = nx$, we can relate the first term in the above bias upper bound to a classical positive linear operator as follows. For any continuous real function $F\in\mathscr{C}[0,1]$, the n-th order *Szász-Mirakyan operator* (Szasz, 1950) $\boldsymbol{S}_n:\mathscr{C}[0,1]\to\mathscr{C}[0,1]$ maps F to

$$\boldsymbol{S}_n[F,x] := \mathop{\mathbb{E}}_{N\sim\mathrm{Poi}(nx)}\left[F\left(\frac{N}{n}\right)\right] = e^{-nx}\sum_{i=0}^{\infty}\frac{(nx)^i}{i!}F\left(\frac{i}{n}\right),$$

where we set $F(x)=F(1)$ for $x>1$ iff $F(x)$ is not defined for $x>1$. The following lemma shows that $\boldsymbol{S}_n[F,x]$ closely approximates F whenever F is a Lipschitz function.

Lemma 10 *For any $n\in\mathbb{N}$, $x\in[0,1]$, and 1-Lipschitz function $F\in\mathscr{C}[0,1]$,*

- *the function $\boldsymbol{S}_n[F,x]$ is also 1-Lipschitz;*
- *we have the following point-wise error bound: $|\boldsymbol{S}_n[F,x]-F(x)|\le\sqrt{x/n}$.*

Proof We present two proofs – one is probabilistic, and the other is analytical. Let $N\sim\mathrm{Poi}(nx)$ and $N'\sim\mathrm{Poi}(nx')$ with $x'>x$. Couple N' and N such that $N'-N\sim\mathrm{Poi}(n(x'-x))$. Then,

$$\begin{aligned}|\boldsymbol{S}_n[F,x']-\boldsymbol{S}_n[F,x]| &= \left|\mathbb{E}[F\left(\frac{N'}{n}\right)-F\left(\frac{N}{n}\right)]\right| \\ &\le \mathbb{E}\left|F\left(\frac{N'}{n}\right)-F\left(\frac{N}{n}\right)\right| \\ &\le \frac{\mathbb{E}|N'-N|}{n} \\ &= |x'-x|.\end{aligned}$$

This establishes the first bullet. For the second, note that

$$\begin{aligned}|\boldsymbol{S}_n(F,x) - F(x)| &\le \mathbb{E}\left|F\left(\frac{N}{n}\right) - F(x)\right| \\ &\le \mathbb{E}\left|\frac{N}{n} - x\right| \\ &\le \sqrt{\frac{x}{n}},\end{aligned}$$

where the last step follows by the Cauchy-Schwarz inequality.

Next, we provide an analytical proof, which shines light on the proof of Lemma 15. We again begin with the first bullet and show that the operator preserves Lipschitzness. Denote $x' := x + s$ and assume that $s \ge 0$. Then we can express $\boldsymbol{S}_n[F, x']$ as

$$\begin{aligned}\boldsymbol{S}_n[F,x'] &\overset{(a)}{=} e^{-nx'}\sum_{i=0}^{\infty}\frac{(nx')^i}{i!}F\left(\frac{i}{n}\right) \\ &\overset{(b)}{=} e^{-n(x+s)}\sum_{i=0}^{\infty}\frac{(n(x+s))^i}{i!}F\left(\frac{i}{n}\right) \\ &\overset{(c)}{=} e^{-n(x+s)}\sum_{i=0}^{\infty}\frac{1}{i!}\sum_{j=0}^{i}\binom{i}{j}(nx)^j(ns)^{i-j}F\left(\frac{i}{n}\right) \\ &\overset{(d)}{=} e^{-n(x+s)}\sum_{j=0}^{\infty}\frac{(nx)^j}{j!}\sum_{i=j}^{\infty}\frac{j!}{i!}\binom{i}{j}(ns)^{i-j}F\left(\frac{i}{n}\right) \\ &\overset{(e)}{=} e^{-n(x+s)}\sum_{j=0}^{\infty}\frac{(nx)^j}{j!}\sum_{i=j}^{\infty}\frac{(ns)^{i-j}}{(i-j)!}F\left(\frac{i}{n}\right) \\ &\overset{(f)}{=} e^{-nx}\sum_{j=0}^{\infty}\frac{(nx)^j}{j!}\left(e^{-ns}\sum_{i=j}^{\infty}\frac{(ns)^{i-j}}{(i-j)!}F\left(\frac{i}{n}\right)\right),\end{aligned}$$

where (a) and (b) follow by the definitions of S_n and x', respectively; (c) follows by the binomial theorem; (d) follows by re-ordering the summation operators; (e) follows by $\binom{i}{j} = i!/(j!(i-j)!)$; (f) follows by factorizing out e^{-ns}.

Therefore, the difference between $\boldsymbol{S}_n[F, x']$ and $\boldsymbol{S}_n[F, x]$ satisfies

$$\begin{aligned}
\left|\boldsymbol{S}_n[F, x'] - \boldsymbol{S}_n[F, x]\right| &\overset{(a)}{=} \left| e^{-nx'} \sum_{i=0}^{\infty} \frac{(nx')^i}{i!} F\left(\frac{i}{n}\right) - e^{-nx} \sum_{i=0}^{\infty} \frac{(nx)^i}{i!} F\left(\frac{i}{n}\right) \right| \\
&\overset{(b)}{\le} \left| e^{-nx} \sum_{j=0}^{\infty} \frac{(nx)^j}{j!} \left(e^{-ns} \sum_{i=j}^{\infty} \frac{(ns)^{i-j}}{(i-j)!} F\left(\frac{i}{n}\right) - F\left(\frac{j}{n}\right) \right) \right| \\
&\overset{(c)}{=} \left| e^{-nx} \sum_{j=0}^{\infty} \frac{(nx)^j}{j!} \left(e^{-ns} \sum_{i=j}^{\infty} \frac{(ns)^{i-j}}{(i-j)!} F\left(\frac{i}{n}\right) \right.\right. \\
&\qquad \left.\left. - e^{-ns} \sum_{i=j}^{\infty} \frac{(ns)^{i-j}}{(i-j)!} F\left(\frac{j}{n}\right) \right) \right| \\
&\overset{(d)}{\le} e^{-nx} \sum_{j=0}^{\infty} \frac{(nx)^j}{j!} \left(e^{-ns} \sum_{i=j}^{\infty} \frac{(ns)^{i-j}}{(i-j)!} \left| F\left(\frac{i}{n}\right) - F\left(\frac{j}{n}\right) \right| \right) \\
&\overset{(e)}{\le} e^{-nx} \sum_{j=0}^{\infty} \frac{(nx)^j}{j!} \left(e^{-ns} \sum_{i=j}^{\infty} \frac{(ns)^{i-j}}{(i-j)!} \frac{|i-j|}{n} \right) \\
&\overset{(f)}{=} e^{-nx} \sum_{j=0}^{\infty} \frac{(nx)^j}{j!} \cdot s \overset{(g)}{=} s,
\end{aligned}$$

where (a) follows by the definition of S_n; (b) follows by the equality we just established; (c) follows by $1 = e^{-ns} \sum_{i=j}^{\infty} (ns)^{i-j}/(i-j)!$; (d) follows by the triangle inequality; (e) follows by the Lipschitz condition on F; (f) follows by the expectation formula of Poisson random variables; (g) follows by $e^{-nx} \sum_{j=0}^{\infty} (nx)^j/j! = 1$.

The proof is complete by noting that $s = x' - x$. Next we establish the second claim, whose proof closely follows that of Proposition 4.9 in Bustamante (2017).

$$|\boldsymbol{S}_n[F, x] - F(x)| \overset{(a)}{\le} \boldsymbol{S}_n(|F(t) - F(x)|, x) \overset{(b)}{\le} \boldsymbol{S}_n(|t - x|, x) \overset{(c)}{\le} \boldsymbol{S}_n((t-x)^2, x)^{\frac{1}{2}} \overset{(d)}{\le} \sqrt{\frac{x}{n}},$$

where (a) follows by the triangle inequality; (b) follows by the Lipschitzness of F; (c) follows by Lemma 4.1 in Bustamante (2017); and (d) follows by $\mathrm{Var}_{X \sim \mathrm{Poi}(\lambda)}(X) = \lambda$ or direct evaluation. ■

As a corollary, we obtain tight upper bounds on the difference between $\boldsymbol{S}_n[F, x]$ and $\boldsymbol{S}_{na}[F, x]$.

Corollary 4 *For any $n \in \mathbb{N}$, $x \in [0, 1]$, $a \ge 2.5$, and 1-Lipschitz function $F \in \mathscr{C}[0, 1]$,*

$$|\boldsymbol{S}_n[F, x] - \boldsymbol{S}_{na}[F, x]| \le \sqrt{\frac{8x}{3n}}.$$

Using this corollary, we bound the quantity of interest as follows.

$$\left| \mathop{\mathbb{E}}_{X' \sim \mathrm{Poi}(\lambda)} [\mathbb{1}_{X' > \tau}] \left(\mathop{\mathbb{E}}_{Y \sim \mathrm{Poi}(a\lambda)} \left[f\left(\frac{Y}{an}\right) \right] - \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(\lambda)} \left[f\left(\frac{X}{n}\right) \right] \right) \right| \le \frac{\lambda}{n} \sqrt{\frac{8}{3\tau}}.$$

where the first inequality follows by $\mathbb{E}_{X'\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X'>\tau}] = \Pr(\mathrm{Poi}(\lambda) > \tau) \le \sqrt{\Pr(\mathrm{Poi}(\lambda) > \tau)}$ and the corollary; the second follows by applying Markov's inequality, i.e., $\Pr(\mathrm{Poi}(\lambda) > \tau) \le \lambda/\tau$.

It remains to bound the second term on the right-hand side of the aforementioned bias upper bound, for which we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,X'\sim\mathrm{Poi}(\lambda)}\left|\mathbb{1}_{X'\le\tau}\mathbb{1}_{X>2r\tau}\cdot f\left(\frac{X}{n}\right)\right| &\overset{(a)}{=} \mathop{\mathbb{E}}_{X'\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X'\le\tau}]\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(\lambda)}\left[\left|\mathbb{1}_{X>2r\tau}\cdot f\left(\frac{X}{n}\right)\right|\right]\\
&\overset{(b)}{\le} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X\le\tau}]\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(\lambda)}\left[\mathbb{1}_{X>2r\tau}\cdot\frac{X}{n}\right]\\
&\overset{(c)}{=} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(\lambda)}[\mathbb{1}_{X\le\tau}]\cdot\frac{1}{n}\sum_{j>2r\tau}e^{-\lambda}\frac{\lambda^j}{j!}\cdot j\\
&\overset{(d)}{=} \frac{\lambda}{n}\Pr\left(\mathrm{Poi}(\lambda)\le\tau\right)\cdot\Pr\left(\mathrm{Poi}(\lambda)\ge 2r\tau\right)\\
&\overset{(e)}{\le} \exp\left(-\frac{2}{5}r\right)\frac{\lambda}{n},
\end{aligned}$$

where (a) follows by the fact that X and X' are independent; (b) follows by the regularity and Lipschitz conditions on f; (c) follows by expanding the expectation; (d) follows by $\sum_{j>2r\tau}(e^{-\lambda}\lambda^j/j!)\cdot j = \lambda\Pr\left(\mathrm{Poi}(\lambda)\ge 2r\tau\right)$; (e) follows by the following generalization of Lemma 9.

Lemma 11 *For any $\lambda,\tau\ge 0$ and $b>1$,*

$$\Pr(\mathrm{Poi}(\lambda)\le\tau)\cdot\Pr(\mathrm{Poi}(\lambda)\ge b\tau)\le\exp\left(-\left(\frac{(c(b)-1)^2}{2c(b)}+\frac{3(b-c(b))^2}{2(b+2c(b))}\right)\tau\right),$$

where for

$$t(b) := \left(-64-528b^2-8742b^4-1331b^6+54\sqrt{5}\sqrt{64b^4+528b^6+5097b^8+1331b^{10}}\right)^{1/3},$$

$$\begin{aligned}
c(b) := &-\frac{b}{4}+\frac{1}{2}\sqrt{\left\{\frac{b^2}{4}+\frac{1}{15}\left(4+11b^2\right)+\frac{\left(4+11b^2\right)^2}{30t(b)}+\frac{t(b)}{30}\right\}}\\
&+\frac{1}{2}\sqrt{\left\{\frac{b^2}{2}+\frac{2}{15}\left(4+11b^2\right)-\frac{\left(4+11b^2\right)^2}{30t(b)}-\frac{t(b)}{30}\right.}\\
&+\left(\frac{16b}{5}-b^3+\frac{2}{5}b\left(-4-11b^2\right)\right)\Big/\left(4\sqrt{\left[\frac{b^2}{4}+\frac{1}{15}\left(4+11b^2\right)\right.}\right.\\
&\left.\left.\left.+\frac{\left(4+11b^2\right)^2}{30t(b)}+\frac{t(b)}{30}\right]\right)\right\}.
\end{aligned}$$

We postpone the proof of this lemma to Appendix M.

Bounding the bias of the combined estimator By our previous constructions, we naturally estimate $\boldsymbol{S}_m[f,x] = \boldsymbol{S}_m[f,\lambda/n] = P_S(f,\lambda)+P_L(f,\lambda)$ by the unbiased estimator of

$$H(f,\lambda) := H_S(f,\lambda)+H_L(f,\lambda),$$

which can be constructed using $X, X' \sim \mathrm{Poi}(\lambda)$ (see the next section). Denote the resulting estimator by $\hat{H}(f, X, X')$. The absolute bias of this estimator in estimating $H(f, \lambda)$ is at most

$$|P_S(f,\lambda) - H_S(f,\lambda)| + |P_L(f,\lambda) - H_L(f,\lambda)| \le \frac{\lambda}{n} \cdot \frac{5}{2\sqrt{\tau}} = \frac{5x}{2\sqrt{\tau}}.$$

Appendix G. From $d = 1$ to $d = 2$ via a Linear-Operator View

By the derivations in the last section, utilizing the samples $X, X' \sim \mathrm{Poi}(nx)$, we can employ the following estimator to estimate $\boldsymbol{S}_m[f, x]$:

$$\hat{H}_{\boldsymbol{v}}(f, X, X') := \mathbb{1}_{X' \le \tau} \cdot \left(\sum_{u=0}^{\tilde{u}} h^u_{X-u} \cdot f\left(\frac{u}{m}\right) \right) + (1 - \mathbb{1}_{X' \le \tau} \mathbb{1}_{X \le 2r\tau}) \cdot f\left(\frac{X}{n}\right).$$

where $m := an$ and $\boldsymbol{v} := (a, n, \tilde{u}, \tau, r)$ is the vector of parameters. Similar to $\boldsymbol{S}_m$, the expectation of this estimator is **a linear operator** over $\mathscr{C}[0,1]$ as well:

$$\begin{aligned}\boldsymbol{L}_{\boldsymbol{v}}[f, x] := \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx)}[\mathbb{1}_{X \le \tau}] \sum_{u=0}^{\tilde{u}} h^r_a(u, nx) \cdot f\left(\frac{u}{m}\right) \\ + \sum_{v \ge 0} \sum_{w \ge 0} h_1(v, nx) h_1(w, nx)(1 - \mathbb{1}_{v \le \tau} \mathbb{1}_{w \le 2r\tau}) \cdot f\left(\frac{w}{n}\right).\end{aligned}$$

Given this definition, we summarize the results obtained in the last section as follows.

Lemma 12 *For any parameters $a \ge 2.5$, $\tau \ge 1/3$, $\tilde{u} \ge 2.5a\tau$, $r \ge 5(\tilde{u}+1) \vee 10(a-1)$, and function $f \in \mathscr{C}[0,1]$ that is c-Lipschitz,*

$$|\boldsymbol{L}_{\boldsymbol{v}}[f, x] - \boldsymbol{S}_n[f, x]| \le \frac{5c}{2\sqrt{\tau}} \cdot x,$$

To facilitate the subsequent discussions, denote

$$\begin{aligned}|\boldsymbol{L}_{\boldsymbol{v}}|\,[f, x] := \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx)}[\mathbb{1}_{X \le \tau}] \sum_{u=0}^{\tilde{u}} |h^r_a(u, nx)| \cdot f\left(\frac{u}{m}\right) \\ + \sum_{v \ge 0} \sum_{w \ge 0} h_1(v, nx) h_1(w, nx)(1 - \mathbb{1}_{v \le \tau} \mathbb{1}_{w \le 2r\tau}) \cdot f\left(\frac{w}{n}\right),\end{aligned}$$

which is a **positive linear operator** over $\mathscr{C}[0,1]$. The triangle inequality yields a simple relation between $\boldsymbol{L}_{\boldsymbol{v}}$ and $|\boldsymbol{L}_{\boldsymbol{v}}|$:

$$|\boldsymbol{L}_{\boldsymbol{v}}[f, x]| \le |\boldsymbol{L}_{\boldsymbol{v}}|\,[|f|, x].$$

In the special case where f being the *identity function* I, we can further bound the magnitude of $|\boldsymbol{L}_{\boldsymbol{v}}|\,[I, x]$ and $\boldsymbol{S}_n[I, x]$ via the following lemma.

Lemma 13 *For any $x \in [0,1]$ and $n \ge 1$,*

$$\boldsymbol{S}_n[I, x] = I(x) = x,$$

and if in addition, $a \ge 2.5$, $\tau \ge 1/3$, $\tilde{u} \ge 2.5a\tau$, and $r \ge 5(\tilde{u}+1) \vee 10(a-1)$, then

$$|\boldsymbol{L}_{\boldsymbol{v}}[I, x]| \le |\boldsymbol{L}_{\boldsymbol{v}}|\,[I, x] \le \left(1 + \frac{2}{3} e^{-r/4}\right) \cdot x.$$

Proof We begin by establishing the first claim.

$$\boldsymbol{S}_n[I,x] = e^{-nx}\sum_{i=0}^{\infty}\frac{(nx)^i}{i!}I\left(\frac{i}{n}\right) = x\cdot e^{-nx}\sum_{i=1}^{\infty}\frac{(nx)^{i-1}}{(i-1)!} = x.$$

As for the second claim, the inequality $|\boldsymbol{L}_{\boldsymbol{v}}[I,x]| \le |\boldsymbol{L}_{\boldsymbol{v}}|[I,x]$ follows by the triangle inequality, and the inequality $|\boldsymbol{L}_{\boldsymbol{v}}|[I,x] \le \left(1+\frac{2}{3}e^{-r/4}\right)\cdot x$ follows by

$$\begin{aligned}
|\boldsymbol{L}_{\boldsymbol{v}}|[I,x] &\overset{(a)}{=} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X\le\tau}]\sum_{u=0}^{\tilde{u}}|h_a^r(u,nx)|\cdot\frac{u}{m} \\
&\quad + \sum_{v\ge 0}\sum_{w\ge 0}h_1(v,nx)h_1(w,nx)(1-\mathbb{1}_{v\le\tau}\mathbb{1}_{w\le 2r\tau})\cdot\frac{w}{n} \\
&\overset{(b)}{\le} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X\le\tau}]\sum_{u=0}^{\tilde{u}}|h_a^r(u,nx)|\cdot\frac{u}{m} + \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X>\tau}]\cdot x + e^{-2r/5}x \\
&\overset{(c)}{\le} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X\le\tau}]\sum_{u=0}^{\tilde{u}}\left(\Pr(\mathrm{Poi}(anx)=u)+e^{-nx}(nx)\cdot e^{-r/3}\right)\frac{u}{m} \\
&\quad + \left(\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X>\tau}]+e^{-2r/5}\right)x \\
&\overset{(d)}{\le} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X\le\tau}]\left(\sum_{u=0}^{\infty}\Pr(\mathrm{Poi}(anx)=u)\cdot\frac{u}{m}+\sum_{u=0}^{\tilde{u}}\frac{ux}{a}\cdot e^{-r/3}\right) \\
&\quad + \left(\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X>\tau}]+e^{-2r/5}\right)x \\
&\overset{(e)}{=} \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X\le\tau}]\left(1+\frac{\tilde{u}(\tilde{u}+1)}{2a}\cdot e^{-r/3}\right)\cdot x + \left(\mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(nx)}[\mathbb{1}_{X>\tau}]+e^{-2r/5}\right)x \\
&\overset{(f)}{\le} \left(1+\frac{\tilde{u}(\tilde{u}+1)}{2a}\cdot e^{-r/3}+e^{-2r/5}\right)\cdot x \\
&\overset{(g)}{\le} \left(1+\frac{2}{3}e^{-r/4}\right)\cdot x,
\end{aligned}$$

where (a) follows by the definition of $|\boldsymbol{L}_{\boldsymbol{v}}|[I,\cdot]$; (b) follows by $h_1(v,nx) = e^{-nx}(nx)^v/v!$, the equality $1-\mathbb{1}_{v\le\tau}\mathbb{1}_{w\le 2r\tau} = \mathbb{1}_{v<\tau}+\mathbb{1}_{v\le\tau}\mathbb{1}_{w>2r\tau}$, and Lemma 11 and the reasoning before it; (c) follows by Lemma 8 and grouping the last two terms together; (d) follows by $e^{-nx}\le 1$ and $m = na$; (e) follows by $\mathbb{E}[\mathrm{Poi}(nax)] = nax = mx$ and $\sum_{u=0}^{\tilde{u}}u = (\tilde{u}+1)\tilde{u}/2$; (f) follows by $\mathbb{E}[\mathbb{1}_A]+\mathbb{E}[\mathbb{1}_{\bar{A}}] = 1$ and re-organizing the terms; (g) follows by the conditions $r \ge 5(\tilde{u}+1)\vee 10(a-1)$ and $a\ge 2.5$, and the inequality $\frac{1}{5}(\frac{r}{5})^2e^{-r/3}+e^{-2r/5}\le\frac{2}{3}e^{-r/4}, \forall r\ge 3.5$. ■

To facilitate the consecutive discussions, we define the "inverse" of the *Szász-Mirakyan operator* $\boldsymbol{S}_n$ as $\boldsymbol{S}_n^{-1}$, which satisfies, for any $F\in\mathscr{C}[0,1]$, that

$$\boldsymbol{S}_n^{-1}[\boldsymbol{S}_n[F,\cdot],x] = F(x).$$

For our purpose, this "inverse operator" is well-defined in the following sense.

Lemma 14 *For any* 1*-Lipschitz functions* $F, G \in \mathscr{C}[0,1]$*, if for all* $n \in \mathbb{N}$ *and all* $x \in [0,1]$*,*

$$\boldsymbol{S}_n[F,x] = \boldsymbol{S}_n[G,x],$$

then the two functions must be identical, i.e., $F = G$*.*

The correctness of this lemma follows from Lemma 10.

Next we consider the $d = 2$ version of our estimation problem. Specifically, given a real function $f(x_1, x_2) \in \mathscr{C}[0,1]^2$ that is 1-Lipschitz with respect to both x_1 and x_2, we want to approximate

$$f_3(x_1,x_2) = \mathop{\mathbb{E}}_{X_i \sim \mathrm{Poi}(m_i x_i)} \left[f\left(\frac{X_1}{m_1}, \frac{X_2}{m_2}\right)\right] = \boldsymbol{S}_{m_2}\left[\boldsymbol{S}_{m_1}\left[f(\cdot,x_2),x_1\right](x_1,\cdot),x_2\right].$$

To make the expression more manageable, we denote

$$f_2(x_1,x_2) = \boldsymbol{S}_{m_2}^{-1}\left[f_3(x_1,\cdot),x_2\right] = \mathop{\mathbb{E}}_{X_1 \sim \mathrm{Poi}(m_1 x_1)} \left[f\left(\frac{X_1}{m_1}, x_2\right)\right].$$

Then, for $\boldsymbol{v}_2 := (a_2, n_2, \tilde{u}_2, \tau_2, r_2)$, we can write $\boldsymbol{L}_{\boldsymbol{v}_2}[f_2(x_1,\cdot),x_2]$ as

$$\begin{aligned}
\boldsymbol{L}_{\boldsymbol{v}_2}[f_2(x_1,\cdot),x_2] &= \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(n_2 x_2)}[\mathbb{1}_{X \le \tau_2}] \sum_{u=0}^{\tilde{u}_2} h_{a_2}^{r_2}(u, n_2 x_2) f_2\left(x_1, \frac{u}{m_2}\right) \\
&\quad + \sum_{v \ge 0}\sum_{w \ge 0} h_1(v, n_2 x_2) h_1(w, n_2 x_2)(1 - \mathbb{1}_{v \le \tau_2}\mathbb{1}_{w \le 2r_2\tau_2}) f_2\left(x_1, \frac{w}{n_2}\right) \\
&= \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(n_2 x_2)}[\mathbb{1}_{X \le \tau_2}] \sum_{u=0}^{\tilde{u}_2} h_{a_2}^{r_2}(u, n_2 x_2) \mathop{\mathbb{E}}_{X_1 \sim \mathrm{Poi}(m_1 x_1)} \left[f\left(\frac{X_1}{m_1}, \frac{u}{m_2}\right)\right] \\
&\quad + \sum_{v \ge 0}\sum_{w \ge 0} h_1(v, n_2 x_2) h_1(w, n_2 x_2)(1 - \mathbb{1}_{v \le \tau_2}\mathbb{1}_{w \le 2r_2\tau_2}) \mathop{\mathbb{E}}_{X_1 \sim \mathrm{Poi}(m_1 x_1)} \left[f\left(\frac{X_1}{m_1}, \frac{w}{n_2}\right)\right].
\end{aligned}$$

Move the inner expectation outside and define

$$\begin{aligned}
f_1(x_1,x_2) &:= \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(n_2 x_2)}[\mathbb{1}_{X \le \tau_2}] \sum_{u=0}^{\tilde{u}_2} h_{a_2}^{r_2}(u, n_2 x_2) f\left(x_1, \frac{u}{m_2}\right) \\
&\quad + \sum_{v \ge 0}\sum_{w \ge 0} h_1(v, n_2 x_2) h_1(w, n_2 x_2)(1 - \mathbb{1}_{v \le \tau_2}\mathbb{1}_{w \le 2r_2\tau_2}) f\left(x_1, \frac{w}{n_2}\right).
\end{aligned}$$

Then we establish the identity

$$\boldsymbol{L}_{\boldsymbol{v}_2}[f_2(x_1,\cdot),x_2] = \mathop{\mathbb{E}}_{X_1 \sim \mathrm{Poi}(m_1)} \left[f_1\left(\frac{X_1}{m_1}, x_2\right)\right] = \boldsymbol{S}_{m_1}[f_1(\cdot,x_2),x_1].$$

According to Lemma 10, function f_2 is also 1-Lipschitz with respect to both of its arguments. The key observation is that for properly chosen hyper-parameters, the function

$$f_1(\cdot,x_2) := \boldsymbol{S}_{m_1}^{-1}[\boldsymbol{L}_{\boldsymbol{v}_2}[f_2(x_1,\cdot),x_2](\cdot,x_2),\cdot] = \boldsymbol{L}_{\boldsymbol{v}_2}[f(x_1,\cdot),x_2](\cdot,x_2)$$

is almost 1-Lipschitz for any $x_2 \in [0,1]$.

Lemma 15 *For any function $f : [0,1]^2 \to \mathbb{R}$ that it 1-Lipschitz with respect to both of its arguments, and any parameter vector $\boldsymbol{v} := (a, n, \tilde{u}, \tau, r)$ satisfying the conditions in Lemma 12, the function*

$$\tilde{f}(\cdot, x_2) := L_{\boldsymbol{v}}[f(x_1, \cdot), x_2](\cdot, x_2)$$

is $\left(1 + \frac{4}{3}e^{-r/4}\right)$-Lipschitz for any $x_2 \in [0,1]$.

Proof For any two real values $x_1 \neq x_1' \in [0,1]$ and $x_2 \in [0,1]$,

$$\begin{aligned}
\frac{|\tilde{f}(x_1, x_2) - \tilde{f}(x_1', x_2)|}{|x_1 - x_1'|} &\overset{(a)}{\leq} \frac{1}{|x_1 - x_1'|} \Bigg| \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} \mathbb{1}_{X \leq \tau} \sum_{u=0}^{\tilde{u}} h_a^r(u, nx_2) \times \\
&\left(f\left(x_1, \frac{u}{m}\right) - f\left(x_1', \frac{u}{m}\right)\right) + \sum_{v \geq 0} \sum_{w \geq 0} h_1(v, n_2 x_2) h_1(w, n_2 x_2) \times \\
&(1 - \mathbb{1}_{v \leq \tau_2} \mathbb{1}_{w \leq 2r_2\tau_2}) \left(f\left(x_1, \frac{w}{n}\right) - f\left(x_1', \frac{w}{n}\right)\right) \Bigg| \\
&\overset{(b)}{\leq} \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} [\mathbb{1}_{X \leq \tau}] \sum_{u=0}^{\tilde{u}} |h_a^r(u, nx_2)| \\
&+ \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} [\mathbb{1}_{X > \tau}] + \exp\left(-\frac{2}{5}r\right) \\
&\overset{(c)}{\leq} \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} [\mathbb{1}_{X \leq \tau}] \sum_{u=0}^{\tilde{u}} \left(\Pr(\mathrm{Poi}(mx_2) = u) + e^{-r/3}\right) \\
&+ \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} [\mathbb{1}_{X > \tau}] + \exp\left(-\frac{2}{5}r\right) \\
&\overset{(d)}{\leq} 1 + \mathop{\mathbb{E}}_{X \sim \mathrm{Poi}(nx_2)} [\mathbb{1}_{X \leq \tau}] \left(\frac{\tilde{u}(\tilde{u}+1)}{2e^{r/3}}\right) + \exp\left(-\frac{2}{5}r\right) \\
&\overset{(e)}{\leq} 1 + \frac{\tilde{u}(\tilde{u}+1)}{2e^{r/3}} + e^{-2r/5} \\
&\overset{(f)}{\leq} 1 + \frac{r}{5}\left(\frac{r}{5} - 1\right) \frac{1}{2e^{r/3}} + e^{-2r/5} \\
&\overset{(g)}{\leq} 1 + \frac{4}{3e^{r/4}},
\end{aligned}$$

where (a) follows by the definition of $L_{\boldsymbol{v}}$; (b) follows by the Lipschitzness of f; (c) follows by Lemma 8 and reorganizing the terms; (d) follows by the definition of h_a and the law of total probability; (e) follows by $\mathbb{E}[\mathbb{1}_A] = \Pr(A) \leq 1$; (f) follows by the assumption that $r \geq 5(\tilde{u}+1)$; (g) follows by the inequality $\frac{r}{5}\left(\frac{r}{5} - 1\right)\frac{1}{2e^{r/3}} + e^{-2r/5} \leq \frac{4}{3e^{r/4}}$ for all $r \geq -1$. ■

For brevity, we conclude our discussion about the $d = 2$ case with this lemma, showing that our estimator essentially preserves the Lipschitzness of the function to approximate. In the next section, we leverage this lemma to bound the bias of our estimator for general additive properties.

Appendix H. From $d = 2$ to $d > 2$ By Induction

For any $x \in [0,1]^d$ and $i \in [d]$, denote

$$(x;y)_i := (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d).$$

For any function $f : \mathbb{R}^d \to \mathbb{R}$, define

$$f_{d+1}((x)) := \mathop{\mathbb{E}}_{N_i \sim \mathrm{Poi}(m_i x_i)} \left[f\left(\left(\frac{N_i}{m_i} \right)_{i=1}^d \right) \right]$$

and

$$f_d((x)) := \boldsymbol{S}_{m_d}^{-1} \left[f_{d+1}\left((x;\cdot)_d \right), x_d \right].$$

For $i \in [d-1]$, we define a sequence of functions inductively.

$$f_i((x)) := \boldsymbol{S}_{m_i}^{-1} \left[\boldsymbol{L}_{\boldsymbol{v}_{i+1}} \left[f_{i+1}((x;\cdot)_{i+1}), x_{i+1} \right] \left((x;\cdot)_i \right), x_i \right],$$

and

$$f_0((x)) := \boldsymbol{L}_{\boldsymbol{v}_1} \left[f_1((x;\cdot)_1), x_1 \right].$$

By Lemma 15, for any $i \in [d]$, the function f_i is Lipschitz with a Lipschitz constant

$$C_i := \prod_{t=i+1}^{d} \left(1 + \frac{4}{3e^{r_t/4}} \right).$$

We approximate f_{d+1} with f_0. Next we bound the approximation error

$$\begin{aligned}
|f_0((x)) - f_{d+1}((x))| \le & |f_0((x)) - \boldsymbol{S}_{m_1}\left[f_1((x;\cdot)_1), x_1 \right]| + |\boldsymbol{S}_{m_d}\left[f_d((x;\cdot)_d), x_d \right] - f_{d+1}((x))| \\
& + \sum_{i=1}^{d-1} \left| \boldsymbol{S}_{m_i}\left[f_i((x;\cdot)_i), x_i \right] - \boldsymbol{S}_{m_{i+1}}\left[f_{i+1}((x;\cdot)_{i+1}), x_{i+1} \right] \right| \\
= & |\boldsymbol{L}_{\boldsymbol{v}_1}\left[f_1((x;\cdot)_1), x_1 \right] - \boldsymbol{S}_{m_1}\left[f_1((x;\cdot)_1), x_1 \right]| \\
& + \sum_{i=1}^{d-1} \left| \boldsymbol{L}_{\boldsymbol{v}_{i+1}}\left[f_{i+1}((x;\cdot)_{i+1}), x_{i+1} \right] - \boldsymbol{S}_{m_{i+1}}\left[f_{i+1}((x;\cdot)_{i+1}), x_{i+1} \right] \right| \\
\le & \sum_{i=1}^{d} \frac{5C_i}{2\sqrt{\tau_i}} \cdot x_i = \sum_{i=1}^{d} \frac{5x_i}{2\sqrt{\tau_i}} \cdot \prod_{t=i+1}^{d} \left(1 + \frac{4}{3e^{r_t/4}} \right) \\
\le & \sum_{i=1}^{d} \frac{4}{\sqrt{\tau_i}} x_i,
\end{aligned}$$

where we have made the simple assumption that $3d \le \min_i e^{r_i/4}$ and $5e^{4/9} < 8$. Note that we can replace the multiplicative factor of 4 with 3.9, which is used in Appendix J for clean expressions.

Bounding the bias of the general estimator Recall that the quantity of interest is

$$\sum_{j\in[k]} \mathop{\mathbb{E}}_{M_{i,j}\sim\mathrm{Poi}(m_i p_i(j))} \left[f_j \left(\left(\frac{M_{i,j}}{m_i} \right)_{i=1}^{d} \right) \right] = \sum_{j\in[k]} f_{j,d+1}(p(j)).$$

Following the above reasoning, we approximate this quantity by an *unbiased estimator* of

$$f_0(p) := \sum_{j\in[k]} f_{j,0}(p(j)).$$

The exact form of this estimator is postponed to Section I.2, where we define this estimator and analyze its variance. Given the last inequality $|f_0((x)) - f_{d+1}((x))| \le \sum_{i=1}^{d} 4x_i/\sqrt{\tau_i}$, the bias of this estimator is at most

$$\sum_{j\in[k]} \sum_{i=1}^{d} \frac{4}{\sqrt{\tau_i}} p_i(j) = \sum_{i=1}^{d} \frac{4}{\sqrt{\tau_i}} \sum_{j\in[k]} p_i(j) = \sum_{i=1}^{d} \frac{4}{\sqrt{\tau_i}}.$$

Appendix I. Variance Analysis

I.1. Basic case: $d = 1$

Utilizing the samples $X, X' \sim \mathrm{Poi}(nx)$, we employ the following estimator to estimate $\boldsymbol{S}_m[f, x]$:

$$\hat{H}_{\boldsymbol{v}}(f, X, X') = \mathbb{1}_{X'\le\tau} \cdot \left(\sum_{u=0}^{\tilde{u}} h^u_{X-u} \cdot f\left(\frac{u}{m}\right) \right) + (1 - \mathbb{1}_{X'\le\tau}\mathbb{1}_{X\le 2r\tau}) \cdot f\left(\frac{X}{n}\right).$$

where we recall that $h^u_s = 0, \forall s < 0$, and

$$h^u_s = a^u(1-a)^s \binom{s+u}{u} \Pr\left(\mathrm{Poi}(r) > s + 2u\right), \forall s \ge 0.$$

The variance of this estimator satisfies

$$\begin{aligned}
\mathrm{Var}\left(\hat{H}_{\boldsymbol{v}}(f, X, X')\right) &\overset{(a)}{\le} 2\,\mathbb{E}\left(\sum_{u=0}^{\tilde{u}} h^u_{X-u} \cdot f\left(\frac{u}{m}\right) \right)^2 + 4\mathrm{Var}\left(\mathbb{1}_{X'>\tau} \cdot f\left(\frac{X}{n}\right) \right) \\
&\quad + 4\mathrm{Var}\left(\mathbb{1}_{X'\le\tau}\mathbb{1}_{X>2r\tau} \cdot f\left(\frac{X}{n}\right) \right) \\
&\overset{(b)}{\le} 2\,\mathbb{E}\left(\sum_{u=0}^{\tilde{u}} h^u_{X-u} \cdot f\left(\frac{u}{m}\right) \right)^2 + x \cdot \frac{8(\tau+3)}{n} \\
&\overset{(c)}{\le} 2\Pr(\mathrm{Poi}(nx) \ge 1) \cdot \max_t \left(\sum_{u=0}^{\tilde{u}} h^u_{t-u} \cdot f\left(\frac{u}{m}\right) \right)^2 + x \cdot \frac{8(\tau+3)}{n},
\end{aligned}$$

where (a) follows by $\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y)$; (b) follows by Lemma 11 in the supplementary of Hao et al. (2018) and *the next lemma*; (c) follows by $f(0) = 0$ and the linearity of expectation.

Lemma 16 *For any $n \in \mathbb{Z}^+$, $r, \tau, x \geq 0$, and independent random variables $X, X' \sim \mathrm{Poi}(nx)$,*

$$\mathrm{Var}\left(\mathbb{1}_{X' \leq \tau}\mathbb{1}_{X > 2r\tau} \cdot f\left(\frac{X}{n}\right)\right) \leq x \cdot \frac{\tau + 3}{n}.$$

Proof Following the independence assumption,

$$\begin{aligned}
\mathrm{Var}\left(\mathbb{1}_{X' \leq \tau}\mathbb{1}_{X > 2r\tau} \cdot f\left(\frac{X}{n}\right)\right) &\overset{(a)}{=} \mathrm{Var}(\mathbb{1}_{X' \leq \tau})\mathbb{E}\left[\mathbb{1}^2_{X > 2r\tau} \cdot f^2\left(\frac{X}{n}\right)\right] \\
&\quad + (\mathbb{E}[\mathbb{1}_{X' \leq \tau}])^2 \mathrm{Var}\left(\mathbb{1}_{X > 2r\tau} \cdot f\left(\frac{X}{n}\right)\right) \\
&\overset{(b)}{\leq} \mathbb{E}\left[\mathbb{1}_{X' < \tau}\right] \cdot \mathbb{E}\left[f^2\left(\frac{X}{n}\right)\right] + \mathrm{Var}\left(\mathbb{1}_{X > 2r\tau} \cdot f\left(\frac{X}{n}\right)\right) \\
&\overset{(c)}{\leq} x \cdot \frac{\tau + 2}{n} + \frac{1}{2}\mathrm{Var}\left(\mathbb{1}_{X > 2r\tau} f\left(\frac{X}{n}\right) - \mathbb{1}_{X' > 2r\tau} f\left(\frac{X'}{n}\right)\right) \\
&\overset{(d)}{=} x \cdot \frac{\tau + 2}{n} + \frac{1}{2}\,\mathbb{E}\left(\mathbb{1}_{X > 2r\tau} f\left(\frac{X}{n}\right) - \mathbb{1}_{X' > 2r\tau} f\left(\frac{X'}{n}\right)\right)^2 \\
&\overset{(e)}{\leq} x \cdot \frac{\tau + 2}{n} + \frac{1}{2}\,\mathbb{E}\left(\mathbb{1}_{X > 2r\tau} \cdot \frac{X}{n} - \mathbb{1}_{X' > 2r\tau} \cdot \frac{X'}{n}\right)^2 \\
&\overset{(f)}{=} x \cdot \frac{\tau + 2}{n} + \frac{\mathrm{Var}(\mathbb{1}_{X > 2r\tau} \cdot X)}{n^2} \\
&\overset{(g)}{\leq} x \cdot \frac{\tau + 2}{n} + \frac{\mathrm{Var}(X)}{n^2} \\
&\overset{(h)}{=} x \cdot \frac{\tau + 3}{n},
\end{aligned}$$

where (a) follows by $\mathrm{Var}(Y \cdot Z) = \mathrm{Var}(Y)\mathbb{E}[Z^2] + (\mathbb{E}[Y])^2\mathrm{Var}(Z)$ for any independent Y and Z; (b) follows by $\mathrm{Var}(\mathbb{1}_A) = \mathbb{E}[\mathbb{1}_A] \cdot \mathbb{E}[\mathbb{1}_{\bar{A}}]$ and the fact that the expectation of an indicator random variable is ≤ 1; (c) follows by the Lipschitzness and regularity of f, the inequality $\mathbb{E}[X^2] \leq (nx)^2 + nx$, and

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{X' < \tau}\right]\left(x^2 + \frac{x}{n}\right) &\leq x\left(\mathbb{E}\left[\mathbb{1}_{X' < \tau}\right] x + \frac{1}{n}\right) \\
&= \frac{x}{n}\left(e^{-nx}\sum_{j=0}^{\tau}\frac{(nx)^{j+1}}{(j+1)!}(j+1) + 1\right) \\
&\leq x \cdot \frac{\tau + 2}{n};
\end{aligned}$$

(d) follows by the equality $\mathrm{Var}(Z) = \mathbb{E}[(Z - Z')^2]/2$ for any i.i.d. random variables Z and Z'; (e) follows by considering four cases induced by the possible values of the indicator functions; (f) follows by the same reasoning as in (d); (g) follows by $\mathrm{Var}(\min\{C, Z\}) \leq \mathrm{Var}(Z)$ for any fixed C; (h) follows by $\mathrm{Var}(X) = nx$. ■

The following lemma bounds the remaining term on the right-hand side of our variance bound.

Lemma 17 *For any index t and coefficients $h_{\cdot}^{u}$ defined at the beginning of this section,*

$$\left|\sum_{u=0}^{\tilde{u}} h_{t-u}^{u} \cdot f\left(\frac{u}{m}\right)\right| \leq \frac{e^{ar}\tilde{u}}{ma}.$$

Proof By the definition of $h_{\cdot}^{u}$, we obtain

$$\begin{aligned}
\left|\sum_{u=0}^{\tilde{u}\wedge t} h_{t-u}^{u} \cdot g\left(\frac{u}{m}\right)\right| &\overset{(a)}{=} \left|\sum_{u=0}^{\tilde{u}\wedge t} a^{u}(1-a)^{t-u}\binom{t}{u}\Pr\left(\mathrm{Poi}(r) > t+u\right) \cdot g\left(\frac{u}{m}\right)\right| \\
&\overset{(b)}{\leq} \sum_{u=0}^{\tilde{u}\wedge t} a^{u}(a-1)^{t-u}\binom{t}{u}\Pr\left(\mathrm{Poi}(r) > t+u\right) \cdot \left|g\left(\frac{u}{m}\right)\right| \\
&\overset{(c)}{\leq} e^{-r}\frac{(a-1)^{t-u}}{m}\sum_{u=0}^{\tilde{u}\wedge t} u a^{u}\binom{t}{u}\sum_{j>t+u}\frac{r^{j}}{j!} \\
&\overset{(d)}{\leq} \frac{e^{-r}}{m}\sum_{u=1}^{\tilde{u}\wedge t}\frac{u}{a^{u}}\binom{t}{u}\sum_{j>t}\frac{(ar)^{j}}{j!} \\
&\overset{(e)}{\leq} \frac{e^{-r}\tilde{u}}{ma}\sum_{j>t}\left(\sum_{u=1}^{\tilde{u}\wedge t}\frac{1}{a^{u}}\binom{t}{u}\right)\frac{(ar)^{j}}{j!} \\
&\overset{(f)}{\leq} \frac{e^{-r}\tilde{u}}{ma}\sum_{j>t}\left(1+\frac{1}{a}\right)^{t}\frac{(ar)^{j}}{j!} \\
&\overset{(g)}{\leq} \frac{e^{-r}\tilde{u}}{ma}\sum_{j>t}\frac{((a+1)r)^{j}}{j!} \\
&\overset{(h)}{\leq} \frac{e^{-r}\tilde{u}}{ma}e^{(a+1)r} \\
&\overset{(i)}{=} \frac{e^{ar}\tilde{u}}{ma},
\end{aligned}$$

where (a) follows by our construction; (b) follows by the triangle inequality; (c) follows by the regularity and Lipschitzness of g, and re-organizing terms; (d) follows by $(a-1)^{t-u}a^{u}\sum_{j>t+u} r^{j}/j! \leq a^{-u}\sum_{j>t+u}(ra)^{j}/j! \leq a^{-u}\sum_{j>t}(ra)^{j}/j!$; (e) follows by re-ordering the summation operators, and the fact that the summation is over $u \leq \tilde{u}$; (f) follows by the binomial theorem; (g) follows by $(1+1/a)^{t} \leq (1+1/a)^{j}$ for $j > t$; (h) follows by the series expansion of $e^{(a+1)r}$; (i) follows by the simple equality $e^{-r}\cdot e^{(a+1)r} = e^{ar}$. ■

Therefore, the variance of the proposed estimator admits

$$\mathrm{Var}\left(\hat{H}_{\boldsymbol{v}}(f, X, X')\right) \leq 2\Pr(\mathrm{Poi}(nx) \geq 1)\cdot\left(\frac{e^{ar}\tilde{u}}{ma}\right)^{2} + x\cdot\frac{8(\tau+3)}{n} \leq \frac{x}{n}\cdot 2\left(e^{2ar}\tau^{2} + 4\tau + 12\right),$$

where the second step follows by Markov's inequality and an additional assumption $\tilde{u} = 2.5a\tau$. The following lemma summarizes and slightly rephrases the previous results.

Lemma 18 *For any 1-Lipschitz function $f \in \mathscr{C}[0,1]$, $\tau \geq 1$, and $\tilde{u} = 2.5a\tau$,*

$$\mathrm{Var}\left(\hat{H}_{\boldsymbol{v}}(f, X, X')\right) \leq 6e^{2ar}\tau^{2}\cdot\frac{x}{n}.$$

I.2. General Case: $d > 1$

For any $x, y \in [0,1]^d$ and $i \in [d]$, denote

$$(x, y; i) := (x_1, \ldots, x_i, y_{i+1}, \ldots, y_d) \text{ and } (x, y; z, i) := (x_1, \ldots, x_{i-1}, z, y_{i+1}, \ldots, y_d)$$

For random variables $X_i, X_i' \sim \mathrm{Poi}(n_i x_i)$ and any function $f_{(1)} := f \in \mathscr{C}[0,1]^d$, we define a sequence of random functions $f_{(i)} \in \mathscr{C}[0,1]^d$ as

$$f_{(i+1)}((x,y;i)) := \mathbb{1}_{X_i' \le \tau_i} \cdot \sum_{u=0}^{\tilde{u}_i} h^u_{X_i - u} \cdot f_{(i)}\left(\left(x, y; \frac{u}{m_i}, i\right)\right) + (1 - \mathbb{1}_{X_i' \le \tau_i} \mathbb{1}_{X_i' \le 2r_i\tau_i}) \cdot f_{(i)}\left(\left(x, y; \frac{X_i}{n_i}, i\right)\right).$$

This construction provides an unbiased estimator for the desired quantity (see Section H), i.e.,

$$\mathop{\mathbb{E}}_{X_i, X_i' \sim \mathrm{Poi}(n_i x_i)} \left[f_{(d+1)}((x))\right] = f_0((x)).$$

Our objective is to bound the variance of $f_{(d+1)}((x))$. By the law of total variance,

$$\mathrm{Var}\left(f_{(i+1)}((x,y;i))\right) = \mathrm{Var}\left(\mathbb{E}\left[f_{(i+1)}((x,y;i)) \middle| X_i, X_i' \sim \mathrm{Poi}(n_i x_i)\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(f_{(i+1)}((x,y;i)) \middle| X_i, X_i' \sim \mathrm{Poi}(n_i x_i)\right)\right]$$

Fix the vector (y) and view $f((x,y;\cdot,i))$ as a function in $\mathscr{C}[0,1]^{i-1}$. For the first quantity on the right-hand side, the construction of $f_{(i+1)}$ yields

$$\mathbb{E}\left[f_{(i+1)}((x,y;\cdot,i)) \middle| X_i, X_i' \sim \mathrm{Poi}(n_i x_i)\right] = \hat{H}\left(f_0((x,y;\cdot,i)), X_i, X_i'\right)$$

By the derivations in the last section (Section H), fixing (x) and (y), the real function $f_0((x,y;\cdot,i))$ is $5/3$-Lipschitz in its argument. Therefore, by Lemma 18,

$$\mathrm{Var}(\hat{H}_{\boldsymbol{v}_i}\left(f_0((x,y;\cdot,i)), X_i, X_i'\right)) \le \left(6\tau_i^2 e^{2a_i r_i}\right) \frac{x_i}{n_i}.$$

Next we consider the second quantity on the right-hand side. By the linearity of expectation,

$$\mathop{\mathbb{E}}_{X_i, X_i' \sim \mathrm{Poi}(n_i x_i)} \mathrm{Var}\left(f_{(i+1)}((x,y;i)) \middle| X_i, X_i'\right) \le \max_{z_i, z_i' \in \mathbb{N}} \mathrm{Var}\left(f_{(i+1)}((x,y;i)) \middle| (X_i, X_i') = (z_i, z_i')\right).$$

We leverage the recursion relation between $f_{(i+1)}$ and $f_{(i)}$ through the following lemma.

Lemma 19 *For any random variables X_i and real numbers c_i,*

$$\mathrm{Var}\left(\sum_i c_i \cdot X_i\right) \le \left(\sum_i |c_i|\right)^2 \max_j \mathrm{Var}\left(X_j\right).$$

Proof Expanding the left-hand side yields

$$\begin{aligned}\mathrm{Var}\left(\sum_i c_i \cdot X_i\right) &= \sum_i c_i^2 \mathrm{Var}\left(X_i\right) + \sum_{i_1 \neq i_2} c_{i_1} c_{i_2} \mathrm{Cov}(X_{i_1}, X_{i_2}) \\ &\leq \sum_i c_i^2 \mathrm{Var}\left(X_i\right) + \sum_{i_1 \neq i_2} \left|c_{i_1} c_{i_2}\right| \sqrt{\mathrm{Var}\left(X_{i_1}\right) \cdot \mathrm{Var}\left(X_{i_2}\right)} \\ &\leq \left|\sum_i c_i^2 + \sum_{i_1 \neq i_2} \left|c_{i_1} c_{i_2}\right|\right| \cdot \max_j \mathrm{Var}\left(X_j\right) \\ &= \left(\sum_i |c_i|\right)^2 \max_j \mathrm{Var}\left(X_j\right),\end{aligned}$$

where the second step follows by the covariance inequality. ■

This lemma, together with the relation between $f_{(i+1)}$ and $f_{(i)}$, yields that

$$\mathrm{Var}\left(f_{(i+1)}((x, y; i))\middle| X_i = z_i, X_i' = z_i'\right) \leq \left(\sum_{u=0}^{\tilde{u}_i} \left|h_{z_i - u}^u\right| + 1\right)^2 \max_{z \in \mathbb{R}} \mathrm{Var}\left(f_{(i)}\left((x, y; z, i)\right)\right).$$

The next lemma further bounds the value of $\sum_{u=0}^{\tilde{u}_i} \left|h_{z_i-u}^u\right|$.

Lemma 20 *For any $a \geq 2.5$, $r \geq 1$, and $\tilde{u}, z \geq 0$,*

$$\sum_{u=0}^{\tilde{u}} \left|h_{z-u}^u\right| \leq e^{ar} - 1.$$

Proof By the definition of $h_\cdot^u$ and assumption $a \geq 2.5$,

$$\begin{aligned}\sum_{u=0}^{\tilde{u}_i} \left|h_{z-u}^u\right| &\overset{(a)}{=} \sum_{u=0}^{\tilde{u} \wedge z} a^u (a-1)^{z-u} \binom{z}{u} \Pr\left(\mathrm{Poi}(r) > z + u\right) \\ &\overset{(b)}{=} e^{-r} \sum_{u=0}^{\tilde{u} \wedge z} a^u (a-1)^{z-u} \binom{z}{u} \sum_{j > z+u} \frac{r^j}{j!} \\ &\overset{(c)}{\leq} \frac{e^{-r}}{a} \sum_{u=0}^{\tilde{u} \wedge z} \frac{1}{a^u} \binom{z}{u} \sum_{j > z+u} \frac{(ar)^j}{j!} \\ &\overset{(d)}{\leq} \frac{e^{-r}}{a} \sum_{j > z} \frac{(ar)^j}{j!} \left(\sum_{u=0}^{\tilde{u} \wedge z} \frac{1}{a^u} \binom{z}{u}\right) \\ &\overset{(e)}{\leq} \frac{e^{-r}}{a} \sum_{j > z} \frac{(ar)^j}{j!} \left(1 + \frac{1}{a}\right)^z \\ &\overset{(f)}{\leq} \frac{e^{-r}}{a} \sum_{j > z} \frac{((a+1)r)^j}{j!} \\ &\overset{(g)}{\leq} \frac{e^{-r}}{a} e^{(a+1)r} \overset{(h)}{\leq} e^{ar} - 1,\end{aligned}$$

where (a) follows by the definition of $h^u_\cdot$; (b) follows by $\Pr\left(\mathrm{Poi}(r) > z+u\right) = e^{-r}\sum_{j>z+u} r^j/j!$; (c) follows by $a^z \sum_{j>z+u} r^j/j! \le \sum_{j>z+u} (ar)^j/j!$; (d) follows by adding positive terms and reorganizing them; (e) follows by the binomial theorem; (f) follows by $(ar)^j(1+1/a)^z \le ((a+1)r)^j$ for $j > z$; (g) follows by the expansion of e^x; (h) follows by $e^x/2.5 \le e^x - 1, \forall x \ge 2.5$. ■

Assume that $a_i \ge 2.5$ and $r_i \ge 10(a_i - 1)$, then $e^{a_i r_i} > 1$. Consolidating the previous results yields

$$\mathrm{Var}\left(f_{(i+1)}((x,y;i))\right) \le e^{2a_i r_i} \cdot \left(\frac{6\tau_i^2 x_i}{n_i} + \max_{z\in\mathbb{R}} \mathrm{Var}\left(f_{(i)}\left((x,y;z,i)\right)\right)\right).$$

In addition, for the special case of $i = 1$, we note that

$$f_{(2)}((x,y;1)) = \hat{H}_{\boldsymbol{v}_1}\left(f((x,y;\cdot,1)), X_1, X_1'\right),$$

which, together with Lemma 18, implies

$$\max_{z\in\mathbb{R}} \mathrm{Var}\left(f_{(2)}((x,y;z,2))\right) \le \left(6\tau_1^2 e^{2a_1 r_1}\right)\frac{x_1}{n_1}.$$

Mathematical induction combines the last two inequalities and yields that

- For every $i \in [d]$, we have an upper bound b_i on $\mathrm{Var}\left(f_{(i+1)}((x,y;i))\right)$ that depends on (x) through only $x_1, \ldots, x_i$.
- The upper-bound sequence $\{b_i\}_{i=1}^d$ satisfies $b_1 = \left(2.1\tau_1^2 e^{2a_1 r_1}\right) x_1/n_1$ and the recurrent relation
$$b_i \le e^{2a_i r_i}\left(\frac{6\tau_i^2 x_i}{n_i} + b_{i-1}\right),\ \forall i \ge 2.$$

Dividing both sides by $c_i := \prod_{t=1}^i e^{2a_t r_t}$, we can rewrite the recurrent relation as

$$\frac{b_i}{c_i} \le \frac{6\tau_i^2 x_i}{n_i c_{i-1}} + \frac{b_{i-1}}{c_{i-1}}.$$

Note that $b_1/c_1 = 6\tau_1^2 x_1/n_1$. The above inequality implies

$$\frac{b_d}{c_d} \le \sum_{i=2}^d \frac{6\tau_i^2 x_i}{n_i c_{i-1}} + \frac{6\tau_1^2 x_1}{n_1} \iff \mathrm{Var}\left(f_{(d+1)}((x))\right) \le \sum_{i=2}^d \frac{6\tau_i^2 x_i}{n_i}\prod_{t=i}^d e^{2a_t r_t} + \frac{6\tau_1^2 x_1}{n_1}\prod_{t=1}^d e^{2a_t r_t}.$$

For simplicity, we adopt the following variance upper bound.

$$\mathrm{Var}\left(f_{(d+1)}(x)\right) \le \sum_{i=1}^d \frac{6\tau_i^2 x_i}{n_i}\exp\left(2\sum_{t=i}^d a_t r_t\right).$$

Bounding the variance of the general estimator The quantity of interest is

$$\sum_{j\in[k]} \mathop{\mathbb{E}}_{N_{i,j}\sim\mathrm{Poi}(m_i p_i(j))} \left[f_j \left(\left(\frac{N_{i,j}}{m_i} \right)_{i=1}^{d} \right) \right] = \sum_{j\in[k]} f_{j,d+1}(p(j)).$$

Following the above derivations, we approximate this quantity by

$$f_{(d+1)}(p) := \sum_{j\in[k]} f_{j,(d+1)}(p(j)).$$

Due to Poisson sampling, the empirical counts $N_{i,j}$ are mutually independent. The variance of estimator, given the last inequality, is at most

$$\mathrm{Var}\left(f_{(d+1)}(p)\right) = \sum_{j\in[k]} \mathrm{Var}\left(f_{j,(d+1)}(p(j))\right) \le 6\sum_{i=1}^{d} \frac{\tau_i^2}{n_i} \exp\left(2\sum_{t=i}^{d} a_t r_t\right).$$

Appendix J. Special Case: n_i's are Equal

In this section we consider the special case where n_i's are equal. We use the same hyper-parameters for all the distributions and suppress the indices in their expressions, i.e., we write r instead of r_i. Then by the results in Section H and above, our proposed estimator $f_{(d+1)}(p)$ (defined above) admits an absolute bias bound of

$$\left|\mathbb{E}[f_{(d+1)}(p)] - \mathbb{E}[f^E(\{X_i^{m_i}\}_{i=1}^d)]\right| \le \frac{3.9d}{\sqrt{\tau}},$$

and a variance bound of

$$\mathrm{Var}\left(f_{(d+1)}(p)\right) \le 6\frac{\tau^2}{n}\sum_{i=1}^{d} \exp\left(2(d-i+1)ar\right) = \frac{6\tau^2}{1-e^{-2ar}} \cdot \frac{e^{2ard}-1}{n},$$

given that $a \ge 2.5$, $\tau \ge 1$, $\tilde{u} = 2.5a\tau$, $3d \le e^{r/4}$, and $r \ge 5(\tilde{u}+1) \vee 10(a-1)$. For simplicity, we choose $r = 15a\tau$. Then the variance bound vanishes at a rate of $\mathcal{O}_\tau(n^{-1/6})$ if

$$2ard = 30a^2\tau d \le \frac{5}{6}\log n \iff 6a\sqrt{\tau d} \le \sqrt{\log n} \implies \tau = \mathcal{O}(\log n).$$

Through Chebyshev's inequality, we combine these results and establish Theorem 1.

Theorem 1 *For any $a \ge 2.5$, $\tau \ge 1$, if $\frac{2\log d}{\tau} \le 6a \le \sqrt{\frac{\log n}{\tau d}}$,*

$$\Pr\left(\left|f_{(d+1)}(p) - \mathbb{E}\left[f^E(\{X_i^m\}_{i=1}^d)\right]\right| \ge \frac{4d}{\sqrt{\tau}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{n^{1/6}}\right).$$

Appendix K. Inferring High-Dimensional Independence

Let $d \in \mathbb{Z}^+$ be a dimension parameter and let $k := (k_1, \ldots, k_d)$ be a vector of *alphabet sizes*.

Denote by Δ_{k_i} the collection of distributions (marginals) over $[k_i] := \{1, \ldots, k_i\}$. Let $p := (p_1, \ldots, p_d)$ be a tuple of distributions in $\Delta_k^\times := \prod_{i=1}^d \Delta_{k_i}$, and for each $j := (j_1, \ldots, j_d) \in [k] := \prod_{i=1}^d [k_i]$, denote by $p(j) := (p_1(j_1), \ldots, p_d(j_d))$ the vector of the corresponding probabilities. Unlike the previous setting, we denote by Δ_k the collection of distributions over $[k]$ and consider a unknown distribution $\tilde{p} \in \Delta_k$.

Denote by $p^\times \in \Delta_k$ the product distribution of p_i's, i.e., $p^\times(j) = \prod_{i=1}^d p_i(j_i)$. We want to estimate, using independent samples from $\tilde{p}$ and p_i's,

$$\ell_1\left(\tilde{p}, p^\times\right) := \sum_{j \in [k]} \left|\tilde{p}(j) - p^\times(j)\right| = \sum_{j \in [k]} \left|\tilde{p}(j) - \prod_{i=1}^d p_i(j_i)\right|,$$

the ℓ_1 distance between $\tilde{p}$ and the product distribution $p^\times$. This defines the basic problem of tolerant high-dimensional independence testing. Note that the property is generally *not* covered (not additive) by the previous results unless $d=1$ as each $p_i(j_i)$ appears $\prod_{t \neq i} k_t$ times on the right-hand side.

First we show that $\ell_1\left(\tilde{p}, p^\times\right)$ is 1-Lipschitz with respect to each of its arguments if we fix the remaining ones. Clearly, the property is 1-Lipschitz with respect to each $\tilde{p}(j)$. By symmetry, it suffices to consider the Lipschitzness for a particular $p_i(j_i)$, for which the desired result follows by

$$\left|\sum_{j' : j'_i = j_i} \left(\left|\tilde{p}(j') - p_i(j'_i) \prod_{t \neq i} p_t(j'_t)\right| - \left|\tilde{p}(j') - (p_i(j'_i) + z) \prod_{t \neq i} p_t(j'_t)\right|\right)\right| \leq z \sum_{j' : j'_i = j_i} \prod_{t \neq i} p_t(j'_t) = z,$$

for any real number z in the unit interval.

For every $i \in [d]$, we denote the following quantities. Let n_i be a *sampling parameter*, let a_i be an *amplification parameter*, and let $Y_i^{M_i} \sim p_i$ be a sample of size $M_i \sim \mathrm{Poi}(m_i)$ where $m_i := a_i \cdot n_i$. For each $j_i \in [k_i]$, denote by M_{i,j_i} the number of times symbol j_i appearing in $Y_i^{M_i}$. Slightly abusing the notation, we write M_{j_i} instead of M_{i,j_i}. Analogously, for distribution $\tilde{p}$ and any symbol $j \in [k]$, we denote $\tilde{n}, \tilde{a}, \tilde{m}$, $\tilde{Y}^{\tilde{M}}$, and $\tilde{M}_j$ in a similar manner.

For any $(x) \in [0,1]^d$ and $y \in [0,1]$, define a function

$$f((x), y) := \left|y - \prod_{t=1}^d x_i\right| - y - \prod_{t=1}^d x_i$$

and extend it by constants for input values larger than 1. Similar to the formulation in the previous sections, we consider approximating the expected value of the empirical estimator

$$\ell_1^E\left(\left\{Y_i^{M_i}\right\}_{i=1}^d, \tilde{Y}^{\tilde{M}}\right) := \sum_{j \in [k]} f\left(\left(\frac{M_{j_i}}{m_i}\right)_{i=1}^d, \frac{\tilde{M}_j}{\tilde{m}}\right).$$

For each index $i \in [d+1]$, we denote $\tilde{S}_i := S_{m_1} \circ \ldots \circ S_{m_{i-1}}$ and $\tilde{L}_i := L_{\boldsymbol{v}_i} \circ \ldots \circ L_{\boldsymbol{v}_d}$, and write

$$f_i\left((x), y\right) = S_{\tilde{m}} \circ \tilde{S}_i \circ \tilde{L}_i\left[f\right]\left((x), y\right),$$

where the linear operator $S_{\tilde{m}}$ applies to the last coordinate of f, and S_{m_t} and $L_{\boldsymbol{v}_t}$ apply to the t-th coordinate of f, for all $t \in [d]$. Then, we can write the expectation of the empirical estimator as

$$\mathbb{E}\left[\ell_1^E\left(\left\{Y_i^{M_i}\right\}_{i=1}^d, \tilde{Y}^{\tilde{M}}\right)\right] = \tilde{f}(p,\tilde{p}) := \sum_{j\in[k]} f_{d+1}\left(p(j),\tilde{p}(j)\right).$$

Clearly, function f is 1-Lipschitz with respect to each of its arguments. By Lemma 15,

$$f_0((x),y) := \boldsymbol{S}_{\tilde{m}}^{-1}[f_1]((x),y) = \tilde{L}_1\left[f\right]((x),y) = L_{\boldsymbol{v}_1} \circ \ldots \circ L_{\boldsymbol{v}_d}\left[f\right]((x),y)$$

is $e^{4/9}$-Lipschitz with respect to y under the conditions specified therein. For some vector $\tilde{v}$ of parameters to be determined later, we denote $f_{-1}((x),y) := L_{\tilde{v}}\left[f_0((x),\cdot),y\right]$, and approximate the expectation of the empirical estimator using an unbiased estimator (Section K.2) of

$$\hat{f}(p,\tilde{p}) := \sum_{j\in[k]} f_{-1}\left(p(j),\tilde{p}(j)\right).$$

K.1. Bias Analysis

We naturally bound the bias of this estimator by

$$\begin{aligned}\left|\hat{f}(p,\tilde{p}) - \tilde{f}(p,\tilde{p})\right| &\le \left|\sum_{j\in[k]}\left(f_{-1}\left(p(j),\tilde{p}(j)\right) - f_{d+1}\left(p(j),\tilde{p}(j)\right)\right)\right| \\ &\le \sum_{j\in[k]}\left|f_{-1}\left(p(j),\tilde{p}(j)\right) - f_1\left(p(j),\tilde{p}(j)\right)\right| \\ &\quad + \sum_{i=1}^{d}\left|\sum_{j\in[k]} f_i\left(p(j),\tilde{p}(j)\right) - f_{i+1}\left(p(j),\tilde{p}(j)\right)\right|,\end{aligned}$$

where both steps follow by the triangle inequality. Note that $f_1 = S_{\tilde{m}}[f_0]$ and $f_{-1} = L_{\boldsymbol{v}}[f_0]$. Hence, for $\boldsymbol{v}$ satisfying the conditions in Lemma 12, the first term on the right-hand side is at most

$$\sum_{j\in[k]}\left|S_{\tilde{m}}[f_0]\left(p(j),\tilde{p}(j)\right) - L_{\tilde{v}}[f_0]\left(p(j),\tilde{p}(j)\right)\right| \overset{(a)}{\le} \sum_{j\in[k]}\frac{5e^{4/9}}{2\sqrt{\tilde{\tau}}}\cdot\tilde{p}(j) \overset{(b)}{<} \frac{4}{\sqrt{\tilde{\tau}}},$$

where (a) follows by Lemma 12 and the fact that $f_0((x),y)$ is $e^{4/9}$-Lipschitz with respect to y; and (b) follows by $\sum_{j\in[k]}\tilde{p}(j) = 1$ and $5e^{4/9} < 8$.

For the second term on the right-hand side, by symmetry, we need to consider only

$$\begin{aligned}
&\left|\sum_{j\in[k]} f_i\left(p(j),\tilde{p}(j)\right) - f_{i+1}\left(p(j),\tilde{p}(j)\right)\right| \\
=&\left|\sum_{j\in[k]} S_{\tilde{m}}\circ\tilde{S}_i\circ\tilde{L}_i\left[f\right]\left((x),y\right) - S_{\tilde{m}}\circ\tilde{S}_{i+1}\circ\tilde{L}_{i+1}\left[f\right]\left((x),y\right)\right| \\
=&\left|\sum_{j\in[k]} \left(L_{\boldsymbol{v}_i} - S_{m_i}\right)\circ S_{\tilde{m}}\circ\tilde{S}_i\circ\tilde{L}_{i+1}\left[f\right]\left(p(j),\tilde{p}(j)\right)\right| \\
\leq&\sum_{t\in[k_i]}\left|\left(L_{\boldsymbol{v}_i} - S_{m_i}\right)\left[\sum_{j:j_i=t} S_{\tilde{m}}\circ\tilde{S}_i\circ\tilde{L}_{i+1}\left[f\right]\left(\left(p(j);\cdot,i\right),\tilde{p}(j)\right),p_i(t)\right]\right|.
\end{aligned}$$

Fixing p and $\tilde{p}$, for any $i\in[d]$ and $t\in[k_i]$, we define $g_{i,t}$ as a real function satisfying

$$g_{i,t}(z;p,\tilde{p}) = \sum_{j:j_i=t} S_{\tilde{m}}\circ\tilde{S}_i\circ\tilde{L}_{i+1}\left[f\right]\left(\left(p(j);z,i\right),\tilde{p}(j)\right).$$

Then, we can write the previous upper bound as $\sum_{t\in[k_i]}\left|\left(L_{\boldsymbol{v}_i} - S_{m_i}\right)\left[g_{i,t}(\cdot\,;p,\tilde{p}),p_i(t)\right]\right|$. The following lemma shows that $g_{i,t}(z;p,\tilde{p})$ is $e^{4/9}$-Lipschitz when viewed as a function of z.

Lemma 21 *$\forall i\in[d]$, $t\in[k_i]$, p, and $\tilde{p}$, if $3d\leq e^{r_i/4}$, $\forall i$, then $g_{i,t}(z;p,\tilde{p})$ is $e^{4/9}$-Lipschitz in z.*

Proof Let $\pi:\mathbb{R}^d\to\mathbb{R}$ be a function satisfying $\pi((x)) = \prod_{i=1}^d x_i, \forall(x)\in[0,1]^d$ and extend it by constants for values larger than 1. For notational convenience, write $f_y((x)) := f((x),y)$. Then,

for any $z_1, z_2 \in [0,1]$, the function values differ by

$$
\begin{aligned}
B_{i,t}(z_1, z_2; p, \tilde{p}) &:= |g_{i,t}(z_1; p, \tilde{p}) - g_{i,t}(z_2; p, \tilde{p})| \\
&\overset{(a)}{=} \left| \sum_{j:j_i=t} S_{\tilde{m}} \circ \tilde{S}_i \circ \tilde{L}_{i+1} \left[((x), y) \to (f_y((x; z_1)_i) - f_y((x; z_2)_i)) \right] ((p(j)), \tilde{p}(j)) \right| \\
&\overset{(b)}{\leq} \left| \sum_{j:j_i=t} S_{\tilde{m}} \circ \tilde{S}_i \circ \left| \tilde{L}_{i+1} \right| \left[((x), y) \to |f_y((x; z_1)_i) - f_y((x; z_2)_i)| \right] ((p(j)), \tilde{p}(j)) \right| \\
&\overset{(c)}{\leq} \left| \sum_{j:j_i=t} S_{\tilde{m}} \circ \tilde{S}_i \circ \left| \tilde{L}_{i+1} \right| \left[((x), y) \to |\pi((x; z_1)_i) - \pi((x; z_2)_i)| \right] ((p(j)), \tilde{p}(j)) \right| \\
&\overset{(d)}{=} |z_1 - z_2| \cdot \left| \sum_{j:j_i=t} \tilde{S}_i \circ \left| \tilde{L}_{i+1} \right| \left[(x) \to \prod_{i' \neq i} (x_{i'} \wedge 1) \right] (p(j)) \right| \\
&\overset{(e)}{\leq} |z_1 - z_2| \cdot \left| \sum_{j:j_i=t} \tilde{S}_i \circ \left| \tilde{L}_{i+1} \right| \left[(x) \to \prod_{i' \neq i} x_{i'} \right] (p(j)) \right| \\
&\overset{(f)}{=} |z_1 - z_2| \cdot \left| \sum_{j:j_i=t} \left| L_{\boldsymbol{v}_{i+1}} \right| \circ \ldots \circ |L_{\boldsymbol{v}_d}| \circ \right. \\
&\qquad \left. S_{m_1} \circ \ldots \circ S_{m_{i-1}} \left[(x) \to \prod_{i' \neq i} x_{i'} \right] (p(j)) \right| \\
&\overset{(g)}{=} |z_1 - z_2| \cdot \left| \sum_{j:j_i=t} \left(\prod_{i'<i} p_i(j_{i'}) \right) \left| L_{\boldsymbol{v}_{i+1}} \right| \circ \ldots \circ |L_{\boldsymbol{v}_d}| \left[(x) \to \prod_{i'>i} x_{i'} \right] (p(j)) \right| \\
&\overset{(h)}{=} |z_1 - z_2| \cdot \left| \sum_{j_{i'} \in [k_{i'}], \forall i'>i} \prod_{i''>i} \left| L_{\boldsymbol{v}_{i''+1}} \right| [I] \left(p_{i''+1}(j_{i''+1}) \right) \right| \\
&\overset{(i)}{\leq} |z_1 - z_2| \cdot \left| \sum_{j_{i'} \in [k_{i'}], \forall i'>i} \prod_{i''>i} \left(\left(1 + \frac{2}{3} e^{-r_{i''}/4} \right) \cdot p_{i''}(j_{i''}) \right) \right| \\
&\overset{(j)}{=} |z_1 - z_2| \cdot \prod_{i'>i} \left(1 + \frac{2}{3} e^{-r_{i'}/4} \right) \sum_{j_{i'} \in [k_{i'}], \forall i'>i} \prod_{i''>i} p_{i''}(j_{i''}) \\
&\overset{(k)}{=} |z_1 - z_2| \cdot \prod_{i'>i} \left(1 + \frac{2}{3} e^{-r_{i'}/4} \right) \overset{(l)}{\leq} |z_1 - z_2| \cdot e^{4/9},
\end{aligned}
$$

where (a) follows by the linearity of linear operators; (b) follows by the triangle inequality and definition of $|\tilde{L}_{i+1}|$; (c) follows by the inequality $|f_y((x; z_1)_i) - f_y((x; z_2)_i)| \leq \pi((x; z_1)_i) - \pi((x; z_2)_i)$; (d) follows by factorizing out $|z_1 - z_2|$; (e) follows by $\prod_{i' \neq i}(x_{i'} \wedge 1) \leq \prod_{i' \neq i} x_{i'}$ where $a \wedge b := \min\{a, b\}$; (f) follows by the definition of $\tilde{S}_i$; (g) follows by the fact that the operator S_{m_t}

preserves the identity function over $[0,1]$ (Lemma 15); (h) follows by summing up $\prod_{i'<i} p_i(j_{i'})$ over all the possible indices; (i) again follows by Lemma 15; (j) follows by re-organizing terms; (k) follows by summing up $\prod_{i''>i} p_{i''}(j_{i''})$; and (l) follows by $3d \le e^{r_i/4}, \forall i$. ■

Combining this with Lemma 12 and the $4/\sqrt{\tilde{\tau}}$ bound above yields

$$\left|\hat{f}(p,\tilde{p}) - \tilde{f}(p,\tilde{p})\right| \le \frac{4}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]}\sum_{t\in[k_i]} \left|(L_{\boldsymbol{v}_i} - S_{m_i})\left[g_{i,t}(\cdot\,;p,\tilde{p}), p_i(t)\right]\right| \le \frac{4}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{4}{\sqrt{\tau_i}}.$$

K.2. Deviation Analysis

Fix $y \ge 0$ and recall that $f_y((x)) = f((x),y)$. Beginning with the function f_y, we denote a sequence of functions over $D := [0,1]^d \times \mathbb{N}^{2d}$ as follows. For every $g \in \mathscr{R}(D)$ and $i \in [d]$, let $\hat{H}_i : \mathscr{R}(D) \to \mathscr{R}(D)$ be a linear operator satisfying

$$\hat{H}_i[g]((x),(z)) := \mathbb{1}_{z_{2i}\le\tau_i} \sum_{u_i=0}^{\tilde{u}_i} h^{u_i}_{z_{2i-1}-u_i}\, g\left(\left(x; \frac{u_i}{m_i}\right)_i, (z)\right) + \left(1 - \mathbb{1}_{z_{2i}\le\tau_i}\mathbb{1}_{z_{2i-1}\le 5\tau_i}\right) g\left(\left(x; \frac{z_{2i-1}}{n_i}\right)_i, (z)\right).$$

For any vectors $(x) \in [0,1]^d$ and $(z) \in \mathbb{N}^{2d}$, let

$$f_{(y,0)}((x),(z)) := f_y((x),(z)) := f_y((x)),$$

and for every $i \in [d]$, let

$$f_{(y,i)} := \hat{H}_i\left[f_{(y,i-1)}\right].$$

It is clear from the construction that $f_{(y,d)}((x),(z))$ corresponds to a particular instantiation of our estimator. In the following, we show that this function is not sensitive to changes in its input values. We bound the mean deviation probability of our estimator through the well-known McDiarmid's inequality (McDiarmid, 1989; Hao and Orlitsky, 2019a), which we state below for completeness.

Lemma 22 *Let $Y_1, \ldots, Y_m$ be independent random variables taking values in ranges $R_1, \ldots, R_m$, and let $F : R_1 \times \ldots \times R_m \to C$ with the property that if one freezes all but the w^{th} coordinate of $F(y_1, \ldots, y_m)$ for some $1 \le w \le m$, then F fluctuates only by most $c_w > 0$, thus*

$$|F(y_1, \ldots, y_{w-1}, y_w, y_{w+1}, \ldots, y_m) - F(y_1, \ldots, y_{w-1}, y'_w, y_{w+1}, \ldots, y_m)| \le c_w$$

for all $y_j \in R_j$ and $y'_w \in R_w$ for $1 \le j \le m$. Then for any $\lambda > 0$, one has

$$\Pr(|F(Y) - \mathbb{E}[F(Y)]| \ge \lambda\sigma) \le C\exp(-c\lambda^2)$$

for some absolute constants $C, c > 0$, where $\sigma^2 := \sum_{j=1}^m c_j^2$.

Lemma 23 *If the corresponding hyper-parameters of $f_{(\cdot,d)}$ satisfies the conditions in Lemma 12,*

$$\left|f_{(y,d)}((x),(z;z_1+1)_1) - f_{(y,d)}((x),(z))\right| \le \frac{a_1}{n_1}\left(\prod_{i=2}^d \frac{z_i}{n_i}\right) e^{1.5\sum_{i=1}^d a_i r_i}.$$

The same inequality holds if we replace $(z;z_1+1)_1$ by $(z;z_{d+1}+1)_{d+1}$.

Proof For simplicity, suppress y in the sub-script. Increasing z_1 by 1 changes the value of $f_{(d)}$ by

$$
\begin{aligned}
S_f((x),(z)) &:= |f_{(d)}((x),(z;z_1+1)_1) - f_{(d)}((x),(z))| \\
&\overset{(a)}{=} \left|\hat{H}_d \circ \ldots \circ \hat{H}_2[f_{(1)}]((x),(z;z_1+1)_1) - \hat{H}_d \circ \ldots \circ \hat{H}_2[f_{(1)}]((x),(z))\right| \\
&\overset{(b)}{=} \left|\hat{H}_d \circ \ldots \circ \hat{H}_2\left[(x',z') \to \left(f_{(1)}((x'),(z';z_1'+1)_1) - f_{(1)}((x'),(z'))\right)\right]((x),(z))\right| \\
&\overset{(c)}{\leq} \left|\hat{H}_d\right| \circ \ldots \circ \left|\hat{H}_2\right|\left[(x',z') \to \left|f_{(1)}((x'),(z';z_1'+1)_1) - f_{(1)}((x'),(z'))\right|\right]((x),(z)) \\
&\overset{(d)}{\leq} \left|\hat{H}_d\right| \circ \ldots \circ \left|\hat{H}_2\right|\left[(x',z') \to \sum_{u_1=0}^{\tilde{u}_1} \left|h^{u_1}_{z_1'+1-u_1} - h^{u_1}_{z_1'-u_1}\right| \cdot \left|f_y\left(\left(x';\frac{u_1}{m_1}\right)_1\right)\right|\right. \\
&\quad \left. + \left|f_y\left(\left(x';\frac{z_1'+1}{n_1}\right)_1\right) - f_y\left(\left(x';\frac{z_1'}{n_1}\right)_1\right)\right| + \left|f_y\left(\left(x';\frac{5\tau_1}{n_1}\right)_1\right)\right|\right]((x),(z)) \\
&\overset{(e)}{\leq} \left|\hat{H}_d\right| \circ \ldots \circ \left|\hat{H}_2\right|\left[(x',z') \to \frac{1}{n_1}\left(\frac{2}{a_1^2}e^{a_1 r_1}\tilde{u}_1 + 1 + 5\tau_1\right)\prod_{i=2}^{d} x_i'\right]((x),(z)) \\
&\overset{(f)}{=} \frac{1}{n_1}\left(\frac{2}{a_1^2}e^{a_1 r_1}\tilde{u}_1 + 1 + 5\tau_1\right)\prod_{i=2}^{d}\left|\hat{H}_i\right|[((x'),(z')) \to x_i']((x),(z)) \\
&\overset{(g)}{\leq} \frac{1}{n_1}\left(\frac{e^{a_1 r_1}\tilde{u}_1}{2}\right)\prod_{i=2}^{d}\left(\sum_{u_i=0}^{\tilde{u}_i}|h^{u_i}_{z_i-u_i}| \cdot \frac{u_i}{m_i} + \frac{z_i}{n_i}\right) \\
&\overset{(h)}{\leq} \frac{1}{n_1}\left(\frac{e^{a_1 r_1}\tilde{u}_1}{2}\right)\prod_{i=2}^{d}\left(\frac{z_i}{n_i}\left(\left(\sum_{u=0}^{\tilde{u}_i}|h^{u_i}_{z_i-u_i}| \cdot \frac{u_i}{a_i}\right) + 1\right)\right) \\
&\overset{(i)}{\leq} \frac{1}{n_1}\left(\frac{e^{a_1 r_1}\tilde{u}_1}{2}\right)\prod_{i=2}^{d}\frac{z_i}{n_i}\left(\frac{e^{a_i r_i}\tilde{u}_i}{a_i^2} + 1\right) \\
&\overset{(j)}{\leq} \frac{a_1}{n_1}\left(\prod_{i=2}^{d}\frac{z_i}{n_i}\right)\left(\prod_{i=1}^{d}\frac{e^{a_i r_i}\tilde{u}_i}{2a_i}\right) \\
&\overset{(k)}{\leq} \frac{a_1}{n_1}\left(\prod_{i=2}^{d}\frac{z_i}{n_i}\right)e^{1.5\sum_{i=1}^{d} a_i r_i},
\end{aligned}
$$

where (a) follows by the definition of $f_{(d)}$; (b) follows by the linearity of linear operators; (c) follows by the triangle inequality; (d) follows by applying the triangle inequality to each component of $f_{(1)}$; (e) follows by Lemma 20 and the definition of $f_y(\cdot)$; (f) again follows by the linearity of linear operators; (g) follows by the conditions on the hyper-parameters; (h) follows by $h^{u_i}_{z_i-u_i} = 0$ for $z_i = 0$ and $m_i = n_i a_i$; (i) follows by a simple variant of Lemma 20; (j) and (k) follow by re-organizing the terms and some simple inequalities on the hyper-parameters. ■

Write $f_{(y,d)}((x),(z))$ as $f_{(y,d)}((z))$ since it is a constant function with respect to (x). Further fix (z) and view $f_{(y,d)}((z))$ as a function of y. Analogous to the previous construction, we define a linear operator $\tilde{H} : \mathscr{R}([0,1] \times \mathbb{N}^2) \to \mathscr{R}([0,1] \times \mathbb{N}^2)$ as

$$\tilde{H}\left[f_{(\cdot,d)}((z))\right](\tilde{z}_1,\tilde{z}_2) := \hat{H}_{\tilde{v}}\left(f_{(\cdot,d)}((z)),\tilde{z}_1,\tilde{z}_2\right).$$

Given empirical counts $\tilde{N}_j$ and N_{j_i}, and their i.i.d. copies $\tilde{N}'_j$ and N'_{j_i}, our unbiased estimator for the quantity of interest $\hat{f}(p,\tilde{p}) := \sum_{j\in[k]} f_{-1}\left(p(j),\tilde{p}(j)\right)$ is

$$\hat{\boldsymbol{I}}\left((N_{j_i},N'_{j_i})_{i=1}^d,\tilde{N}_j,\tilde{N}'_j\right) := \sum_{j\in[k]} \hat{H}_{\tilde{v}}\left(f_{(\cdot,d)}((N_{j_i},N'_{j_i})_{i=1}^d),\tilde{N}_j,\tilde{N}'_j\right).$$

The next lemma bounds the magnitude of any component of this estimator in terms of its inputs.

Lemma 24 *If the corresponding hyper-parameters of $f_{(\cdot,d)}$ satisfies the conditions in Lemma 12,*

$$\left|\tilde{H}\left[f_{(\cdot,d)}((z))\right](\tilde{z}_1,\tilde{z}_2)\right| \le \tilde{u}\cdot\frac{\tilde{z}_1}{\tilde{n}}e^{\tilde{a}\tilde{r}}\cdot e^{\sum_{i=1}^d a_i r_i},$$

and

$$\left|\tilde{H}\left[f_{(\cdot,d)}((z))\right](\tilde{z}_1+1,\tilde{z}_2)-\tilde{H}\left[f_{(\cdot,d)}((z))\right](\tilde{z}_1,\tilde{z}_2)\right| \le \frac{\tilde{u}}{\tilde{n}}e^{\tilde{a}\tilde{r}}\cdot e^{\sum_{i=1}^d a_i r_i}.$$

Proof The proofs for these upper bounds are similar. Hence we prove only the first upper-bound inequality below. The second inequality follows by the same reasoning and that in Lemma 23.

$$\begin{aligned}
\left|\tilde{H}\left[f_{(\cdot,d)}((z))\right](\tilde{z}_1,\tilde{z}_2)\right| &\overset{(a)}{=} \left|\tilde{H}[\hat{H}_d\circ\ldots\circ\hat{H}_1[f.]((x),(z))](\tilde{z}_1,\tilde{z}_2)\right| \\
&\overset{(b)}{=} \left|\hat{H}_d\circ\ldots\circ\hat{H}_1\left[\tilde{H}[f.](\tilde{z}_1,\tilde{z}_2)\right]((x),(z))\right| \\
&\overset{(c)}{\le} |\hat{H}_d|\circ\ldots\circ|\hat{H}_1|\left[|\tilde{H}|[f.](\tilde{z}_1,\tilde{z}_2)\right]((x),(z)) \\
&\overset{(d)}{\le} |\hat{H}_d|\circ\ldots\circ|\hat{H}_1|\left[\sum_{u=0}^{\tilde{u}}|h^u_{\tilde{z}_1-u}||f_{u/\tilde{m}}|+|f_{\tilde{z}_1/\tilde{n}}|\right]((x),(z)) \\
&\overset{(e)}{\le} |\hat{H}_d|\circ\ldots\circ|\hat{H}_1|\left[\mathbb{1}_{\tilde{z}_1>0}\cdot\frac{e^{\tilde{a}\tilde{r}}\tilde{u}}{\tilde{n}\tilde{a}^2}+\frac{\tilde{z}_1}{\tilde{n}}\right]((x),(z)) \\
&\overset{(f)}{\le} \left(\mathbb{1}_{\tilde{z}_1>0}\cdot\frac{e^{\tilde{a}\tilde{r}}\tilde{u}}{\tilde{n}\tilde{a}^2}+\frac{\tilde{z}_1}{\tilde{n}}\right)|\hat{H}_d|\circ\ldots\circ|\hat{H}_1|\left[((x'),(z'))\to 1\right]((x),(z)) \\
&\overset{(g)}{\le} \left(\mathbb{1}_{\tilde{z}_1>0}\cdot\frac{e^{\tilde{a}\tilde{r}}\tilde{u}}{\tilde{n}\tilde{a}^2}+\frac{\tilde{z}_1}{\tilde{n}}\right)\prod_{i=1}^d e^{a_i r_i} \\
&\overset{(h)}{\le} \tilde{u}\cdot\frac{\tilde{z}_1}{\tilde{n}}e^{\tilde{a}\tilde{r}}\cdot e^{\sum_{i=1}^d a_i r_i},
\end{aligned}$$

where (a) follows by the definition of $f_{(\cdot,d)}$; (b) follows by the linearity of linear operators; (c) follows by the triangle inequality; (d) follows by applying the triangle inequality to each component of $f.$; (e) follows by Lemma 20 and $f_y(\cdot)\le y$; (f) follows by re-organizing the terms; (g) follows by the linearity of linear operators and pairing each operator $|\hat{H}_i|$ with the corresponding argument; (h) follows by simple algebra. ■

Consequently, we can bound the estimator's magnitude by

$$\left|\hat{\boldsymbol{I}}\left((N_{j_i},N'_{j_i})_{i=1}^d,\tilde{N}_j,\tilde{N}'_j\right)\right| \le \sum_{j\in[k]}\tilde{u}\cdot\frac{\tilde{N}_j}{\tilde{n}}e^{\tilde{a}\tilde{r}}\cdot e^{\sum_{i=1}^d a_i r_i} = \tilde{u}\cdot\frac{\tilde{N}}{\tilde{n}}e^{\tilde{a}\tilde{r}}\cdot e^{\sum_{i=1}^d a_i r_i}.$$

Assume that $2\tilde{u} \cdot e^{\tilde{a}\tilde{r}} \cdot e^{\sum_{i=1}^{d} a_i r_i} \leq \tilde{n}$. We modify the estimator slightly by replacing the sample sizes N_i, N_i', $\tilde{N}$, and $\tilde{N}'$ with $\min\{N_i, 2n_i\}$, $\min\{N_i', 2n_i\}$, $\min\{\tilde{N}, 2\tilde{n}\}$, and $\min\{\tilde{N}', 2\tilde{n}\}$, respectively. Note that this step is just used to simply the proof. Below we denote this modified estimator by $\hat{\boldsymbol{I}}^{\star}$.

Suppressing the inputs in expressions, this modification changes the estimator's expectation by

$$
\begin{aligned}
\left|\mathbb{E}[\hat{\boldsymbol{I}}^{\star}] - \mathbb{E}[\hat{\boldsymbol{I}}]\right| &\overset{(a)}{\leq} \left|\mathbb{E}\left[\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}} \,|\tilde{N} \leq 2\tilde{n}\right]\right| + \left|\mathbb{E}\left[(\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}})\mathbb{1}_{\tilde{N}>2\tilde{n}}\right]\right| \\
&\overset{(b)}{\leq} \left|\mathbb{E}\left[\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}} \,|\tilde{N} \leq 2\tilde{n}\right]\right| + \mathbb{E}\left[1 \cdot \mathbb{1}_{\tilde{N}>2\tilde{n}}\right] + \mathbb{E}\left[\frac{\tilde{N}}{2\tilde{n}} \cdot \mathbb{1}_{\tilde{N}>2\tilde{n}}\right] \\
&\overset{(c)}{=} \left|\mathbb{E}\left[\hat{\boldsymbol{I}}^{\star}|\tilde{N} \leq 2\tilde{n}\right] - \mathbb{E}\left[\hat{\boldsymbol{I}} \,|\tilde{N} \leq 2\tilde{n}\right]\right| + \mathbb{E}\left[\mathbb{1}_{\tilde{N}>2\tilde{n}}\right] + \mathbb{E}\left[\frac{\tilde{N}}{2\tilde{n}} \cdot \mathbb{1}_{\tilde{N}>2\tilde{n}}\right] \\
&\overset{(d)}{=} \Bigg|\mathbb{E}\left[(\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}}) \cdot \mathbb{1}_{\exists i, N_i \vee N_i' > 2n_i} | \tilde{N} \leq 2\tilde{n}\right] \\
&\quad + \mathbb{E}\left[(\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}}) \cdot \mathbb{1}_{\forall i, N_i, N_i' \leq 2n_i} | \tilde{N} \leq 2\tilde{n}\right]\Bigg| \\
&\quad + \mathbb{E}\left[\mathbb{1}_{\tilde{N}>2\tilde{n}}\right] + \frac{\tilde{n}}{2\tilde{n}}\,\mathbb{E}\left[\mathbb{1}_{\tilde{N}\geq 2\tilde{n}}\right] \\
&\overset{(e)}{=} \left|\mathbb{E}\left[(\hat{\boldsymbol{I}}^{\star} - \hat{\boldsymbol{I}}) \cdot \mathbb{1}_{\exists i, N_i \vee N_i' > 2n_i} | \tilde{N} \leq 2\tilde{n}\right]\right| + \mathbb{E}\left[\mathbb{1}_{\tilde{N}>2\tilde{n}}\right] + \frac{1}{2}\,\mathbb{E}\left[\mathbb{1}_{\tilde{N}\geq 2\tilde{n}}\right] \\
&\overset{(f)}{\leq} \left|\mathbb{E}\left[2 \cdot \mathbb{1}_{\exists i, N_i \vee N_i' > 2n_i} | \tilde{N} \leq 2\tilde{n}\right]\right| + \frac{3}{2}\,\mathbb{E}\left[\mathbb{1}_{\tilde{N}\geq 2\tilde{n}}\right] \\
&\overset{(g)}{=} 2\,\mathbb{E}\left[\mathbb{1}_{\exists i, N_i \vee N_i' > 2n_i}\right] + \frac{3}{2}\,\mathbb{E}\left[\mathbb{1}_{\tilde{N}\geq 2\tilde{n}}\right] \\
&\overset{(h)}{\leq} 4\sum_{i\in[d]} \mathbb{E}\left[\mathbb{1}_{N_i > 2n_i}\right] + \frac{3}{2}\,\mathbb{E}\left[\mathbb{1}_{\tilde{N}\geq 2\tilde{n}}\right] \\
&\overset{(i)}{\leq} \frac{3}{2}e^{-3\tilde{n}/8} + 4\sum_{i\in[d]} e^{-3n_i/8},
\end{aligned}
$$

where (a) follows by the law of total expectation; (b) follows by the previous bound on the estimators' values and $2\tilde{u} \cdot e^{\tilde{a}\tilde{r}} \cdot e^{\sum_{i=1}^{d} a_i r_i} \leq \tilde{n}$; (c) follows by the linearity of linear operators; (d) follows by the law of total expectation; (e) follows by the fact that if $\forall i, N_i, N_i' \leq 2n_i$, then the two estimators coincide; (f) follows by the same reasoning as in (b); (g) follows by the fact that $\tilde{N}$ and N_i, N_i' are independent; (h) follows by the union bound and the fact that N_i and N_i' are independent; (i) follows by the Chernoff bound for Poisson random variables.

The above argument shows that the modified estimator $\hat{\boldsymbol{I}}^{\star}$ has a bias nearly the same as the original estimator $\hat{\boldsymbol{I}}$. Next, we show that the modified estimator is not sensitive to changes in its input values. Due to independence, the following two sampling schemes are equivalent: (1) choose the sample size, e.g., $n_i^{\star} := \min\{N_i, 2n_i\}$, and draw this many sample points from the corresponding distribution, e.g., $X_i^{n_i^{\star}} \sim p_i \in \Delta_{k_i}$; (2) select the maximum possible sample size, e.g., $2n_i$, draw this many sample points from the corresponding distribution, e.g., $X_i^{2n_i} \sim p_i \in \Delta_{k_i}$, and truncate this sample at a random location picked according to the associated Poisson random

variable, e.g., $N_i \sim \mathrm{Poi}(n_i)$. Hence, $\hat{I}^\star$ is an estimator taking as its inputs the following independent random variables: $\{X_i^{2n_i}\}_{i=1}^d$, $\{(X')_i^{2n_i}\}_{i=1}^d$, $\tilde{X}^{2\tilde{n}}$, $(\tilde{X}')^{2\tilde{n}}$, $\{N_i\}_{i=1}^d$, $\{N_i'\}_{i=1}^d$, $\tilde{N}$, and $\tilde{N}'$.

Note that the total number of independent random variables corresponding to $\tilde{p}$ and each distribution p_i is at most

$$\tilde{n}^\star := 4\tilde{n} + 2 \text{ and } n_i^\star := 4n_i + 2,$$

respectively. By Lemma 23, changing a sample point in $X_{i'}^{2n_{i'}}$, $(X')_{i'}^{2n_{i'}}$, $N_{i'}$, or $N_{i'}'$ changes the estimator's value by at most 2 multiplied by

$$\begin{aligned}
\frac{a_{i'}}{n_{i'}} e^{\tilde{a}\tilde{r}} \cdot e^{1.5\sum_{i=1}^d a_i r_i} \sum_{j\in[k]} \left(\prod_{i\neq i'} \frac{N_{j_i}}{n_i} \right) &= \frac{a_{i'}}{n_{i'}} e^{\tilde{a}\tilde{r}} \cdot e^{1.5\sum_{i=1}^d a_i r_i} \prod_{i\neq i'} \left(\sum_{j_i\in[k_i]} \left(\frac{N_{j_i}}{n_i} \right) \right) \\
&\le \frac{a_{i'}}{n_{i'}} e^{\tilde{a}\tilde{r}} \cdot e^{1.5\sum_{i=1}^d a_i r_i} \prod_{i\neq i'} \left(\frac{2n_i}{n_i} \right) \\
&= \frac{a_{i'}}{n_{i'}} \cdot 2^d \cdot e^{\tilde{a}\tilde{r} + 1.5\sum_{i=1}^d a_i r_i}.
\end{aligned}$$

Let $c_{i'}^\star$ denote the value of the last quantity multiplied by 2. Analogously, by Lemma 24, changing a sample point in $\tilde{X}^{2\tilde{n}}$, $(\tilde{X}')^{2\tilde{n}}$, $\tilde{N}$, or $\tilde{N}'$ changes the estimator's value by at most 2 multiplied by

$$\frac{\tilde{u}}{\tilde{n}} e^{\tilde{a}\tilde{r}} \cdot e^{\sum_{i=1}^d a_i r_i} = \frac{\tilde{u}}{\tilde{n}} e^{\tilde{a}\tilde{r} + \sum_{i=1}^d a_i r_i}.$$

Let $\tilde{c}^\star$ denote the value of the last quantity multiplied by 2. Let $\tilde{S} := \{\tilde{X}^{2\tilde{n}}, (\tilde{X}')^{2\tilde{n}}, \tilde{N}, \tilde{N}'\}$, $S_0 := \emptyset$, and $S_i := \{\{X_t^{2n_t}\}_{t=1}^i, \{(X')_t^{2n_t}\}_{t=1}^i, \{N_t\}_{t=1}^i, \{N_t'\}_{t=1}^i\}$. Combined with McDiarmid's inequality (Lemma 22) and union bound, these deviation upper bounds yield

$$\begin{aligned}
\Pr\left(\left| \hat{\boldsymbol{I}}^\star - \mathbb{E}[\hat{\boldsymbol{I}}^\star] \right| \ge \varepsilon \right) &\le \Pr\left(\left| I - \mathbb{E}_{\tilde{S}}[\hat{\boldsymbol{I}}^\star] \right| + \sum_{i=1}^d \left| \mathbb{E}_{\tilde{S},S_{i-1}}[\hat{\boldsymbol{I}}^\star] - \mathbb{E}_{\tilde{S},S_i}[\hat{\boldsymbol{I}}^\star] \right| \ge \varepsilon \right) \\
&\le \Pr\left(\left| I - \mathbb{E}_{\tilde{S}}[\hat{\boldsymbol{I}}^\star] \right| \ge \varepsilon \right) + \sum_{i=1}^d \Pr\left(\left| \mathbb{E}_{\tilde{S},S_{i-1}}[\hat{\boldsymbol{I}}^\star] - \mathbb{E}_{\tilde{S},S_i}[\hat{\boldsymbol{I}}^\star] \right| \ge \varepsilon \right) \\
&\le 2\exp\left(-\frac{2\varepsilon^2}{\tilde{n}^\star (\tilde{c}^\star)^2} \right) + 2\sum_{i=1}^d \exp\left(-\frac{2\varepsilon^2}{n_i^\star (c_i^\star)^2} \right), \ \forall \varepsilon \ge 0.
\end{aligned}$$

Let $\varepsilon := \frac{1}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{1}{\sqrt{\tau_i}}$. The above probability bound vanishes at a super-linear rate if

$$\tilde{c}^\star \le \frac{\varepsilon}{(\tilde{n}^\star)^{0.6}} \iff 2 \cdot \frac{\tilde{u}}{\tilde{n}} e^{\tilde{a}\tilde{r} + \sum_{i=1}^d a_i r_i} \le \frac{\frac{1}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{1}{\sqrt{\tau_i}}}{(4\tilde{n} + 2)^{0.6}},$$

and for all $i \in [d]$,

$$c_i^\star \le \frac{\varepsilon}{(n_i^\star)^{0.6}} \iff 2 \cdot \frac{a_i}{n_i} \cdot 2^d \cdot e^{\tilde{a}\tilde{r} + 1.5\sum_{i=1}^d a_i r_i} \le \frac{\frac{1}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{1}{\sqrt{\tau_i}}}{(4n_i + 2)^{0.6}}.$$

To summarize, by the triangle inequality, with an error probability of at most

$$\delta^\star := 2\exp\left(-2(4\tilde{n} + 2)^{0.2} \right) + 2\sum_{i=1}^d \exp\left(-2(4n_i + 2)^{0.2} \right),$$

the estimator estimates $\tilde{f}(p,\tilde{p}) = \sum_{j\in[k]} f_{-1}\left(p(j),\tilde{p}(j)\right)$, for any $p,\tilde{p}$, to an additive error of

$$\begin{aligned}\varepsilon^\star := 2\varepsilon + \left|\hat{f}(p,\tilde{p}) - \tilde{f}(p,\tilde{p})\right| \leq & \frac{3}{2}e^{-3\tilde{n}/8} + 4\sum_{i\in[d]} e^{-3n_i/8} + \frac{5}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]}\frac{5}{\sqrt{\tau_i}} \\ \leq & \frac{6}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]}\frac{6}{\sqrt{\tau_i}},\end{aligned}$$

where the last step follows by the assumptions $3e^{-3\tilde{n}/8} \leq 2/\sqrt{\tilde{\tau}}$ and $4e^{-3n_i/8} \leq 1/\sqrt{\tau_i}, \forall i \in [d]$.

Appendix L. Special Case: k_i's Are Not Too Different

In this section we consider the special case where k_i's are not too different. Our objective is to provide a proof sketch for Theorem 4, which we restate below.

Theorem 4 *Assume that $c_1 \log k_0 \leq \log k_i \leq c_2 \log k_0, \forall i \in [d]$, for some k_0 and absolute constants $c_1, c_2 > 0$. Then for any parameters $\varepsilon > 0$ and $d \in \mathbb{Z}^+$, sufficiently large k_0, and distributions $\tilde{p} \in \Delta_{[\boldsymbol{k}]}$ and $p \in \Delta_{\boldsymbol{k}}$, if $\tilde{n} = \Omega\left(\frac{(\prod_i k_i)d^{1/2}}{\sqrt{\log(\prod_i k_i)}\varepsilon^3}\right)$ and $n_i = \Omega\left(\frac{k_i d^{7/2}}{\sqrt{\log k_i}\varepsilon^3}\right), \forall i \in [d]$,*

$$\Pr_{Y^{\tilde{n}}\sim\tilde{p}, X^n\sim p}\left(\left|\hat{f}(Y^{\tilde{n}}, X^n) - \ell_1\left(\tilde{p}, p^\times\right)\right| \geq \varepsilon\right) = \tilde{\mathcal{O}}\left(\frac{1}{\tilde{n}^{1/6}}\right).$$

Let $a_0 := \tilde{a}$, $\tau_0 := \tilde{\tau}$, $\tilde{u}_0 := \tilde{u}$, $r_0 := \tilde{r}$ and $\{a_i, \tau_i, \tilde{u}_i, r_i\}_{i\in[d]}$ be the hyper-parameters of our estimator described in the last section. Let $\lfloor d \rceil := \{0, 1, \ldots, d\}$. Basically, for the results in the last section to hold, we need to ensure that

$$a_i \geq 2.5, \tau_i \geq 1, \tilde{u}_i \geq 2.5 a_i\tau_i, \text{ and } r_i \geq 5(\tilde{u}_i + 1) \vee 10(a_i - 1), \forall i \in \lfloor d \rceil.$$

For our purpose, we can set $\tilde{u}_i = 2.5a_i\tau_i$ and $r_i = 15a_i\tau_i$. Given sampling parameters $\tilde{n}$ and n_i's, let $\tilde{m} := \tilde{a}\tilde{n}$ and $m := (m_i)_{i\in[d]} := (a_i n_i)_{i\in[d]}$ be the amplified sample sizes. Denote by $\breve{p}$ and $\hat{p}$ the empirical distribution of the independent samples $Y^{\tilde{m}} \sim \tilde{p}$ and $X^m \sim p$, respectively.

By the empirical-estimator bias bound in Section E,

$$\left|\ell_1\left(\tilde{p}, p^\times\right) - \mathbb{E}[\ell_1\left(\breve{p}, \hat{p}^\times\right)]\right| \leq \sqrt{\frac{\prod_i k_i}{\tilde{m}}} + \sum_{i=1}^d \sqrt{\frac{k_i}{m_i}}.$$

Hence, for the sampling parameters $\tilde{n}, n_i$'s satisfying the conditions stated in Theorem 4, if we set $\tilde{a} = \Theta(\varepsilon\sqrt{\log(\prod_i k_i)/d})$ and $a_i = \Theta(\varepsilon\sqrt{\log k_i}/d^{3/2})$, where the asymptotic notation hides some sufficiently large absolute constants,

$$\left|\ell_1\left(\tilde{p}, p^\times\right) - \mathbb{E}[\ell_1\left(\breve{p}, \hat{p}^\times\right)]\right| \leq \mathcal{O}(\varepsilon) + \sum_{i=1}^d \mathcal{O}\left(\frac{\varepsilon}{d}\right) = \mathcal{O}(\varepsilon).$$

Next we relate our estimator to this larger-sample-size empirical estimator. Summarizing the results established in the last section, the theorem below characterizes the performance of our estimator.

Theorem 5 *For hyper-parameters satisfying the conditions stated above, if further,*

$$2 \cdot \frac{\tilde{u}}{\tilde{n}} e^{\tilde{a}\tilde{r}+\sum_{i=1}^{d} a_i r_i} \le \frac{\frac{1}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{1}{\sqrt{\tau_i}}}{(4\tilde{n}+2)^{0.6}}$$

and

$$2 \cdot \frac{a_i}{n_i} \cdot 2^d \cdot e^{\tilde{a}\tilde{r}+1.5\sum_{i=1}^{d} a_i r_i} \le \frac{\frac{1}{\sqrt{\tilde{\tau}}} + \sum_{i\in[d]} \frac{1}{\sqrt{\tau_i}}}{(4n_i+2)^{0.6}}, \ \forall i \in [d],$$

then with high probability, the proposed estimator closely approximates the empirical estimator with sampling parameters $\tilde{n}$ and $\tilde{m}_i, i \in [d]$. Specifically,

$$\Pr_{Y^{\tilde{n}}\sim\breve{p},X^n\sim p}\left(\left|\hat{f}(Y^{\tilde{n}},X^n)-\mathbb{E}[\ell_1\left(\breve{p},\hat{p}^{\times}\right)]\right| \ge \frac{6}{\sqrt{\tilde{\tau}}}+\sum_{i\in[d]}\frac{6}{\sqrt{\tau_i}}\right) \le 2e^{-2(4\tilde{n}+2)^{0.2}}+2\sum_{i=1}^{d} e^{-2(4n_i+2)^{0.2}}.$$

Consider parameter settings in the above theorem. For the $\Theta(1/\sqrt{\tilde{\tau}} + \sum_{i\in[d]} 1/\sqrt{\tau_i})$ deviation with respect to $\mathbb{E}[\ell_1\left(\breve{p},\hat{p}^{\times}\right)]$ to be $\Theta(\varepsilon)$, we can simply set $\tilde{\tau} = \Theta(1/\varepsilon^2)$ and $\tau_i = \Theta(d^2/\varepsilon^2)$, where the asymptotic notation hides some sufficiently small absolute constants.

Combined with the conditions stated in Theorem 4, our choice of hyper-parameters leads to

$$\tilde{a}\tilde{r} + 1.5\sum_{i=1}^{d} a_i r_i = 15(\tilde{a}^2\tilde{\tau} + 1.5\sum_{i=1}^{d} a_i^2\tau_i) = \Theta(\log(\prod_i k_i)/d + \sum_{i\in[d]} \log(k_i)/d) = \Theta(\log k_0),$$

where for sufficiently large k_0, we can make the absolute constant hidden in the asymptotic notation arbitrarily small (say, smaller than 0.1) by choosing the aforementioned absolute constants properly.

Note that the above derivation also implies that $\tilde{u}, a_i \le \tilde{a}\tilde{r} \le \Theta(\log k_0)$. Hence for the conditions in Theorem 5 to be satisfied, it suffices to have

$$\Theta\left(\frac{\log k_0}{\tilde{n}^{0.4}} e^{0.1\log k_0}\right) \le \Theta(\varepsilon) \text{ and } \Theta\left(\frac{2^d \log k_0}{n_i^{0.4}} e^{0.1\log k_0}\right) \le \Theta(\varepsilon).$$

By the assumptions in Theorem 4, we have $\tilde{n}, n_i \ge \tilde{\Omega}(k_0)$ for all $i \in [d]$. Choosing sufficiently large k_0 ensures that both inequalities hold and thus completes the proof.

Appendix M. Other Proofs Omitted From the Above

Lemma 9 *For any $\lambda, \tau \ge 0$, $a \ge 2.5$, and $u \ge 2.5a\tau$,*

$$\Pr(\mathrm{Poi}(a\lambda) \ge u) \cdot \Pr(\mathrm{Poi}(\lambda) \le \tau) \le \exp\left(-\frac{3}{8}\tau\right).$$

Proof According to the result in Chung and Lu (2017) (Chapter 2), for any $X \sim \mathrm{Poi}(\mu)$ and $x > 0$,

$$\Pr\left(X \le \mu - x\right) \le e^{-x^2/(2\mu)} \text{ and } \Pr\left(X \ge \mu + x\right) \le e^{-\frac{x^2}{2(\mu+x/3)}}.$$

Following these tail bounds, we split our analysis into three cases according to the value of λ. If $\lambda \leq \tau$, then $a\lambda \leq a\tau$ and it is therefore unlikely to have $\mathrm{Poi}(a\lambda) \geq u \geq 2.5a\tau$. More concretely,

$$\Pr(\mathrm{Poi}(a\lambda) \geq u) \leq \exp\left(-\frac{(2.5a\tau - a\lambda)^2}{2(a\lambda + (2.5a\tau - a\lambda)/3)}\right) \leq \exp\left(-\frac{3}{4}a\tau\right) \leq \exp\left(-\frac{3}{8}\tau\right).$$

If $\lambda \geq 2.5\tau$, then the probability of having $\mathrm{Poi}(\lambda) \leq \tau$ should be small,

$$\Pr(\mathrm{Poi}(\lambda) \leq \tau) \leq \exp\left(-\frac{(\lambda - \tau)^2}{2\lambda}\right) \leq \exp\left(-\frac{1.5^2}{2 \cdot 2.5}\tau\right) \leq \exp\left(-\frac{3}{8}\tau\right).$$

Finally, we address the slightly more complex case of $\lambda \in (\tau, 2.5\tau)$. For notational convenience, we denote $t := \lambda/\tau$, which belongs to $(1, 2.5)$. We will also make use of $a \geq 2.5$.

$$\begin{aligned}
\Pr(\mathrm{Poi}(a\lambda) \geq u) \cdot \Pr(\mathrm{Poi}(\lambda) \leq \tau) &\leq \exp\left(-\frac{(2.5a\tau - a\lambda)^2}{2(a\lambda + (2.5a\tau - a\lambda)/3)}\right) \cdot \exp\left(-\frac{(\lambda - \tau)^2}{2\lambda}\right) \\
&= \exp\left(-\frac{3(2.5 - t)^2}{5 + 4t}a\tau - \frac{(t-1)^2}{2t}\tau\right) \\
&\leq \exp\left(-\inf_{t \in (1, 2.5)}\left(\frac{7.5(2.5-t)^2}{5+4t} + \frac{(t-1)^2}{2t}\right) \cdot \tau\right) \\
&\leq \exp\left(-\frac{3}{8}\tau\right).
\end{aligned}$$

This completes our proof of the lemma. ■

Lemma 11 *For any $\lambda, \tau \geq 0$ and $b > 1$,*

$$\Pr(\mathrm{Poi}(\lambda) \leq \tau) \cdot \Pr(\mathrm{Poi}(\lambda) \geq b\tau) \leq \exp\left(-\left(\frac{(c(b)-1)^2}{2c(b)} + \frac{3(b - c(b))^2}{2(b + 2c(b))}\right)\tau\right),$$

where for

$$t(b) := \left(-64 - 528b^2 - 8742b^4 - 1331b^6 + 54\sqrt{5}\sqrt{64b^4 + 528b^6 + 5097b^8 + 1331b^{10}}\right)^{1/3},$$

$$\begin{aligned}
c(b) := &-\frac{b}{4} + \frac{1}{2}\sqrt{\left\{\frac{b^2}{4} + \frac{1}{15}\left(4 + 11b^2\right) + \frac{\left(4 + 11b^2\right)^2}{30t(b)} + \frac{t(b)}{30}\right\}} \\
&+ \frac{1}{2}\sqrt{\left\{\frac{b^2}{2} + \frac{2}{15}\left(4 + 11b^2\right) - \frac{\left(4 + 11b^2\right)^2}{30t(b)} - \frac{t(b)}{30}\right.} \\
&+ \left(\frac{16b}{5} - b^3 + \frac{2}{5}b\left(-4 - 11b^2\right)\right) \Bigg/ \left(4\sqrt{\left[\frac{b^2}{4} + \frac{1}{15}\left(4 + 11b^2\right)\right.}\right. \\
&\left.\left.\left. + \frac{\left(4 + 11b^2\right)^2}{30t(b)} + \frac{t(b)}{30}\right]\right)\right\}.
\end{aligned}$$

Proof According to the result in Chung and Lu (2017) (Chapter 2), for any $X \sim \mathrm{Poi}(\mu)$ and $x > 0$,

$$\Pr\left(X \le \mu - x\right) \le e^{-x^2/(2\mu)} \text{ and } \Pr\left(X \ge \mu + x\right) \le e^{-\frac{x^2}{2(\mu + x/3)}}.$$

Note that $\Pr(\mathrm{Poi}(\lambda) \le \tau) \cdot \Pr(\mathrm{Poi}(\lambda) \ge b\tau) \le \min\left\{\Pr(\mathrm{Poi}(\lambda) \le \tau),\ \Pr(\mathrm{Poi}(\lambda) \ge b\tau)\right\}$. We consider three cases: $\lambda \le \tau$, $\lambda \ge b\tau$, and $\lambda = c\tau$ where $c \in (1, b)$.

If $\lambda \le \tau$, we bound the desired probability by

$$\Pr(\mathrm{Poi}(\lambda) \ge b\tau) \le \Pr(\mathrm{Poi}(\tau) \ge b\tau) \le \exp\left(-\frac{3(b-1)^2}{2(b+2)}\tau\right).$$

If $\lambda \ge b\tau$, we utilize the upper bound

$$\Pr(\mathrm{Poi}(\lambda) \le \tau) \le \Pr(\mathrm{Poi}(b\tau) \le \tau) \le \exp\left(-\frac{(b-1)^2}{2b}\tau\right).$$

Note that for $b > 1$, we have $3/(b+2) > 1/b$. Hence this upper bound is at least the previous one. Finally, for $\lambda = c\tau$ where $c \in (1, b)$,

$$\begin{aligned}\Pr(\mathrm{Poi}(c\tau) \le \tau) \cdot \Pr(\mathrm{Poi}(c\tau) \ge b\tau) &\le \exp\left(-\frac{(c-1)^2}{2c}\tau\right) \cdot \exp\left(-\frac{3(b-c)^2}{2(b+2c)}\tau\right) \\ &\le \exp\left(-\left(\frac{(c-1)^2}{2c} + \frac{3(b-c)^2}{2(b+2c)}\right)\tau\right)\end{aligned}$$

Minimizing the expression on the right-hand side with respect to c yields the desired inequality. ■

Lemma 12 *For any $\lambda \ge 0$, $b \ge 5$, and independent $X', X \sim \mathrm{Poi}(\lambda)$,*

$$\mathbb{E}[\mathbb{1}_{X' > \tau}\mathbb{1}_{bX' \le X}] \le \exp\left(-\frac{2}{3}\tau\right).$$

Proof Expand the left-hand-side expression,

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{X'>\tau}\mathbb{1}_{bX'\le X}] &\overset{(a)}{=} \sum_{v>\tau}\sum_{w\ge bv}\left(e^{-\lambda}\frac{\lambda^v}{v!}\right)\left(e^{-\lambda}\frac{\lambda^w}{w!}\right)\\
&\overset{(b)}{=} e^{-2\lambda}\sum_{v>\tau}\sum_{w\ge bv}\frac{\lambda^{v+w}}{v!w!}\\
&\overset{(c)}{=} e^{-2\lambda}\sum_{v>\tau}\sum_{w\ge bv}\frac{\lambda^{v+w}}{(v+w)!}\binom{v+w}{v}\\
&\overset{(d)}{=} e^{-2\lambda}\sum_{v+w\ge(b+1)(\tau+1)}\frac{(2\lambda)^{v+w}}{(v+w)!}\sum_{v'=\tau+1}^{\frac{v+w}{b+1}}\left(\frac{1}{2}\right)^{v'}\left(\frac{1}{2}\right)^{v+w-v'}\binom{v+w}{v'}\\
&\overset{(e)}{\le} e^{-2\lambda}\sum_{v+w\ge(b+1)(\tau+1)}\frac{(2\lambda)^{v+w}}{(v+w)!}\Pr\left(\mathrm{bin}(v+w,1/2)\le\frac{v+w}{b+1}\right)\\
&\overset{(f)}{\le} e^{-2\lambda}\sum_{v+w\ge(b+1)(\tau+1)}\frac{(2\lambda)^{v+w}}{(v+w)!}\exp\left(-\left(\frac{b-1}{b+1}\right)^2\frac{v+w}{4}\right)\\
&\overset{(g)}{\le} e^{-2\lambda}\sum_{v+w\ge(b+1)(\tau+1)}\frac{(2\lambda)^{v+w}}{(v+w)!}\exp\left(-\frac{(b-1)^2}{4(b+1)}\tau\right)\\
&\overset{(h)}{\le} \exp\left(-\frac{2}{3}\tau\right)\sum_{v+w\ge(b+1)(\tau+1)}e^{-2\lambda}\frac{(2\lambda)^{v+w}}{(v+w)!}\\
&\overset{(i)}{=} \exp\left(-\frac{2}{3}\tau\right)\cdot\Pr\left(\mathrm{Poi}(2\lambda)\ge(b+1)(\tau+1)\right)\\
&\overset{(j)}{\le} \exp\left(-\frac{2}{3}\tau\right),
\end{aligned}$$

where $(a), (b), (c), (i)$, and (j) follow by simple algebra; (d) follows by re-ordering the summation operators and noticing that the sum is over $w \ge bv$; (e) follows by noting that the last part in the previous expression corresponds to a binomial tail probability; (f) follows by the Chernoff bound for binomial random variables; (g) follows by the fact that the summation is over $(v+w) \ge (b+1)(\tau+1)$; (h) follows by the condition $b \ge 5$. ■

Appendix N. Difference Between Two Empirical Estimators

We consider the difference between the empirical estimates under the Poisson and binomial sampling models. The objective is to show that for a sufficiently smooth function, say $f \in \mathscr{C}[0,1]$ that is 1-Lipschitz, theses two sampling models achieve essentially the same level of performances.

Lemma 13 *For any $n \ge 44$, $x \in [0,1]$, and f that is 1-Lipschitz,*

$$|B_n[f,x] - S_n[f,x]| \le \frac{3x}{n^{1/3}}.$$

Proof Let c be a parameter to be determined later. For any $x \in [c/n, 1]$, we have

$$\begin{aligned}
|B_n[f,x] - S_n[f,x]| &\overset{(a)}{\leq} |B_n[f,x] - f(x)| + |f(x) - S_n[f,x]| \\
&\overset{(b)}{\leq} \sum_{j\geq 0} \left| f\left(\frac{j}{n}\right) - f(x) \right| \mathrm{bin}(n,x,j) + \sum_{j\geq 0} \left| f\left(\frac{j}{n}\right) - f(x) \right| \mathrm{Poi}(nx,j) \\
&\overset{(c)}{=} \frac{1}{n} \sum_{j\geq 0} |j - nx| \,\mathrm{bin}(n,x,j) + \frac{1}{n} \sum_{j\geq 0} |j - nx| \,\mathrm{Poi}(nx,j) \\
&\overset{(d)}{=} \frac{1}{n} \mathop{\mathbb{E}}_{X\sim \mathrm{bin}(n,x)} |X - nx| + \frac{1}{n} \mathop{\mathbb{E}}_{Y\sim \mathrm{Poi}(nx)} |Y - nx| \\
&\overset{(e)}{\leq} \frac{1}{n} \sqrt{\mathrm{Var}\left(\mathrm{bin}(n,x)\right)} + \frac{1}{n} \sqrt{\mathrm{Var}\left(\mathrm{Poi}(nx)\right)} \\
&\overset{(f)}{\leq} \frac{1}{n} \sqrt{nx} + \frac{1}{n} \sqrt{nx} \\
&\overset{(g)}{\leq} x \sqrt{\frac{4}{c}},
\end{aligned}$$

where (a) follows by the triangle inequality; (b) follows by the fact that both operators preserve constant functions; (c) follows by the Lipschitzness of f; (d) follows by the definition of absolute mean deviation; (e) follows by Jensen's inequality; (f) follows by standard results on variances of Poisson and binomial random variables; (g) follows by the condition $x \in [c/n, 1]$.

On the other hand, by the above reasoning, we can assume that $f(0) = 0$. For any $x \in [0, c/n]$,

$$\begin{aligned}
|B_n[f,x] - S_n[f,x]| &\overset{(a)}{\leq} \sum_{j\geq 0} |\mathrm{bin}(n,x,j) - \mathrm{Poi}(nx,j)| \, f\left(\frac{j}{n}\right) \\
&\overset{(b)}{=} \sum_{j\leq 2c} |\mathrm{bin}(n,x,j) - \mathrm{Poi}(nx,j)| \, f\left(\frac{j}{n}\right) \\
&\quad + \sum_{j> 2c} |\mathrm{bin}(n,x,j) - \mathrm{Poi}(nx,j)| \, f\left(\frac{j}{n}\right) \\
&\overset{(c)}{\leq} 2x\left(\frac{2c}{n}\right) + \sum_{j> 2c} |\mathrm{bin}(n,x,j) - \mathrm{Poi}(nx,j)| \left(\frac{j}{n}\right) \\
&\overset{(d)}{\leq} 2x\left(\frac{2c}{n}\right) + \sum_{j> 2c} \mathrm{bin}(n,x,j) \left(\frac{j}{n}\right) + \sum_{j> 2c} \mathrm{Poi}(nx,j) \left(\frac{j}{n}\right) \\
&\overset{(e)}{\leq} 2x\left(\frac{2c}{n}\right) + x \Pr\left(\mathrm{bin}(n-1,x) \geq 2c\right) + x \Pr\left(\mathrm{Poi}(nx) \geq 2c\right) \\
&\overset{(f)}{\leq} 2x\left(\frac{2c}{n}\right) + 2xe^{-0.38c} \\
&\overset{(g)}{\leq} x\left(\frac{5c}{n}\right),
\end{aligned}$$

where (a) follows by the triangle inequality; (b) follows by decomposing the previous summation; (c) follows by noting $f(0) = 0$, f is 1-Lipschitz, and thus $f(j/n) \leq j/n \leq 2c/n$ for $j \leq 0$; (d)

follows by the triangle inequality; (e) follows by

$$\begin{aligned}\sum_{j>2c}\binom{n}{j}x^j(1-x)^{n-j}\frac{j}{n} =& x\sum_{j>2c}\binom{n-1}{j-1}x^{j-1}(1-x)^{(n-1)-(j-1)}\\ \le& x\Pr(\mathrm{bin}(n-1,x)\ge 2c),\end{aligned}$$

and

$$\sum_{j>2c}e^{-nx}\frac{(nx)^j}{j!}\left(\frac{j}{n}\right) = x\sum_{j>2c}e^{-nx}\frac{(nx)^{j-1}}{(j-1)!} = x\Pr(\mathrm{Poi}(nx)\ge 2c);$$

(f) follows by $x\in[0,c/n]$ and the Chernoff bound for Poisson random variables; (g) follows by our choice of c (see below) and $n\ge 44$.

Setting $c=(4/25)^{1/3}n^{2/3}$ and combining the inequalities above imply the desired bound. ■

As an immediate corollary, let f be a d-distribution property satisfying the Lipschitz condition. For any sampling vector $n=(n_1,\ldots,n_d)$ and its Poissonized version $N=(N_1,\ldots,N_d)$ where $N_i\sim\mathrm{Poi}(n_i)$ are independent, the expected values of the corresponding empirical estimator differ by only a fairly small quantity. Specifically, if $n_i\ge 44,\forall i$,

$$\mathop{\mathbb{E}}_{X^n\sim p}\left[\hat{f}^E(X^n)\right]-\mathop{\mathbb{E}}_{Y^N\sim p}\left[\hat{f}^E(Y^N)\right]\le 3\sum_i n_i^{-1/3},$$

where we used the fact that both operators, Bernstein and Szász-Mirakyan, preserve Lipschitzness. See Lemma 10 and Theorem 4.11 in Bustamante (2017) for a proof of this fact.