Canonical Estimation in a Rare-Events Regime

Mesrob I. Ohannessian
Laboratory for Information
and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: mesrob@mit.edu

Vincent Y. F. Tan
Department of
Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706
Email: vtan@wisc.edu

Munther A. Dahleh
Laboratory for Information
and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: dahleh@mit.edu

***Abstract*—We propose a general methodology for performing statistical inference within a 'rare-events regime' that was recently suggested by Wagner, Viswanath and Kulkarni. Our approach allows one to easily establish consistent estimators for a very large class of canonical estimation problems, in a large alphabet setting. These include the problems studied in the original paper, such as entropy and probability estimation, in addition to many other interesting ones. We particularly illustrate this approach by consistently estimating the size of the alphabet and the range of the probabilities. We start by proposing an abstract methodology based on constructing a probability measure with the desired asymptotic properties. We then demonstrate two concrete constructions by casting the Good-Turing estimator as a pseudo-empirical measure, and by using the theory of mixture model estimation.**

I. Introduction

We propose a general methodology for performing statistical inference within the 'rare-events regime' suggested by Wagner, Viswanath and Kulkarni in [1], referred to as WVK hereafter. This regime is a scaling statistical model that strives to capture large alphabet settings, and is characterized by the following notion of a *rare-events source*.

Definition 1. Let $\{(A_n, p_n)\}_{n\in\mathbb{N}}$ be a sequence of pairs where each A_n is an alphabet of finite symbols, and p_n is a probability mass function over A_n. Let X_n be a single sample from p_n, and use it to define a 'shadow' sequence $Z_n = np_n(X_n)$. Let P_n denote the distribution of Z_n. We call $\{(A_n, p_n)\}_{n\in\mathbb{N}}$ a *rare-events source*, if the following conditions hold.

(i) There exists an interval $C = [\check{c}, \hat{c}]$, $0 < \check{c} \leq \hat{c} < \infty$, such that for all $n \in \mathbb{N}$ we have $\frac{\check{c}}{n} \leq p_n(a) \leq \frac{\hat{c}}{n}$ for all $a \in A_n$, or equivalently, P_n is supported on C.

(ii) There exists a random variable Z, such that $Z_n \to Z$ in distribution. Equivalently, there exists a distribution P, such that $P_n \Rightarrow P$ weakly.

To complete the model, we adopt the following sampling scheme. For each n, we draw n independent samples from p_n, and we denote them by $X_{n,1}, \cdots, X_{n,n}$. Using these samples, we are interested in estimating various quantities. WVK consider, among a few others, the following:

- The total (Good-Turing) probabilities of all symbols appearing exactly k times, for each $k \in \mathbb{N}_0$.
- The normalized log-probability of the observed sequence.
- The normalized entropy of the source.
- The relative entropy between the true and empirical distributions.

They also consider two-sequence problems and hypothesis testing, but we focus here on single sequence estimation.

It is striking that many of these quantities can be estimated in such a harsh scaling model, where one cannot hope for the empirical distribution to converge in any traditional sense. However, WVK's estimators have some drawbacks. For example, since they are based on series expansions of the quantities to be estimated, one has to carefully choose the growth rate of partial sums, in order to control convergence properties. More importantly, they are specifically tailored to each individual task. Their consistency is established on a case-by-case basis. What is desirable, and what this paper contributes to, is a methodology for performing more general statistical inference within this regime. Ideally such a framework would allow one to tackle a very large class of canonical estimation problems, and establish consistency more easily.

We may summarize the fundamental ideas behind our approach and the organization of this paper as follows. First, in Section II, we isolate the class of estimation problems that we are interested in as those that asymptotically converge to an integral against P. The quantities studied by WVK fall in this category, and so do other interesting problems such as estimating the size of the alphabet. Other problems, such as estimating the range of the probabilities given by the support interval C, can also be studied in this framework.

Next, in Section III, we propose an abstract solution methodology. At its core, we construct a (random) distribution $\tilde{P}_n$ that converges weakly to P for almost every observation sample. This construction immediately establishes the consistency of natural estimators for the abovementioned quantities, if bounds on C are known. If in addition the rate of the convergence of $\tilde{P}_n$ is established, the framework gives consistent estimators even without bounds on C.

To make this methodology concrete, we build on a core result of WVK that establishes the strong consistency of the Good-Turing estimator. In particular, since the role of the empirical measure is lost, we show in Section IV that we can treat the Good-Turing estimator as a pseudo-empirical measure. Once this is established, we can borrow heavily from the theory of mixture models, where inference is done using

i.i.d. samples, and adapt it to our framework. In Section V, we suggest two approaches for constructing $\tilde{P}_n$: one that is based on maximum likelihood, and another that is based on minimum distance. Both constructions guarantee the almost sure weak convergence of $\tilde{P}_n$ to P, but the latter, under some conditions, also provides the desirable convergence rates.

In Section VI we illustrate the methodology with some examples. In particular, we show how one can consistently estimate the entropy of the source and the probability of the sequence as studied by WVK, but we also propose consistent estimators for the size of the alphabet and for the support interval C.

Notation: Throughout, we use $F(.;.)$ to denote the cumulative distribution of the second argument (which is a probability measure on the real line or on the integers) evaluated at the first argument (which is a point on the real line or an integer).

II. A General Class of Estimation Problems

A. Definitions

Given i.i.d. samples $X_{n,1}, \cdots, X_{n,n}$ from the rare-events source (A_n, p_n), we can pose a host of different estimation problems. Since the alphabet is changing, quantities that depend on explicit symbol labels are not meaningful. Therefore, one ought to only consider estimands that are invariant under re-labeling of the symbols in A_n. In particular, we consider the following class of general estimation problems.

Definition 2. Consider the problem of estimating a sequence $\{Y_n\}_{n\in\mathbb{N}}$ of real-valued random variables using, for every n, the samples $X_{n,1}, \cdots, X_{n,n}$. We call this a *canonical estimation problem* if, for every rare-events source, we have:

$$\mathbf{E}\left[Y_n\right] = \int_C f_n(x)\,\mathrm{d}P_n(x). \tag{1}$$

for some sequence $\{f_n\}$ of continuous real-valued functions on $\mathbb{R}^+$ that converge pointwise to a continuous function f.

It is worth noting that it follows that $\{f_n\}$ and f are also bounded on every closed interval $[a, b]$, $0 < a \leq b < \infty$. Observe that this definition corresponds indeed to estimands that are invariant under re-labeling, in expectation. The following lemma characterizes the limit.

Lemma 1. *For any canonical estimation problem,*

$$\mathbf{E}[Y_n] \to \int_C f(x)\,\mathrm{d}P(x). \tag{2}$$

Proof: Since $P_n \Rightarrow P$, we can apply Skorokhod's theorem ([2], p. 333), to construct a convergent sequence of random variables $\xi_n \to_{\text{a.s.}} \xi$, where $\xi_n \sim P_n$ and $\xi \sim P$. By continuity, it follows that $f_n(\xi_n) \to_{\text{a.s.}} f(\xi)$. By the bounded convergence theorem, we then have $\mathbf{E}[f_n(\xi_n)] \to \mathbf{E}[f(\xi)]$. Since $\mathbf{E}[Y_n] = \mathbf{E}[f_n(\xi_n)]$, and $\int_C f(x)\,\mathrm{d}P(x) = \mathbf{E}[f(\xi)]$, the lemma follows. ∎

It is often more interesting to consider the subclass of canonical problems where there is strong concentration around the mean, and where the Borel-Cantelli lemma applies to give almost sure convergence to the mean.

Definition 3. If a canonical estimation problem further satisfies $|Y_n - \mathbf{E}[Y_n]| \to_{\text{a.s.}} 0$, then call it a *strong canonical problem*. It follows that for strong canonical problems,

$$Y_n \to_{\text{a.s.}} \int_C f(x)\,\mathrm{d}P(x). \tag{3}$$

Using these definitions, a reasonable estimator will at least agree with the limit set forth in Lemma 1. Other modes of convergence may be reasonable, but we would like to exhibit a statistic that almost surely converges to that limit. We make this precise in the following definition.

Definition 4. Given a canonical problem as in Definition 2, a corresponding *estimator* is a sequence $\{\hat{Y}_n\}_{n\in\mathbb{N}}$ such that, for each n, $\hat{Y}_n(a_1, \cdots, a_n)$ is a real-valued function on $(A_n)^n$, to be evaluated on the sample sequence $X_{n,1}, \cdots, X_{n,n}$. A *consistent estimator* is one that obeys

$$\hat{Y}_n(X_{n,1}, \cdots, X_{n,n}) \to_{\text{a.s.}} \int_C f(x)\,\mathrm{d}P(x). \tag{4}$$

For canonical estimation problems that are not necessarily strong, this approach produces an asymptotically unbiased estimator, with asymptotic mean squared error that is no more than the asymptotic variance of the estimand itself. For strong canonical estimation problems, this approach establishes strong consistency, in the sense that the estimator converges to the estimand, almost surely.

B. Examples

To motivate the setting we have just described, we first note that all of the quantities studied by WVK are strong canonical estimation problems. For each quantity, WVK propose an estimator, and individually establish its consistency by showing almost sure convergence to the limit in Lemma 1. In contrast, what we emphasize here is that this can potentially be done *universally* over all strong canonical problems.

To highlight the usefulness of this generalization, we illustrate two important quantities that fall within this framework. We will revisit these in more detail in Section VI. The first quantity is the normalized size of the alphabet: $|A_n|/n$. For this, one can show (see, for example, [3]), that $|A_n|/n = \int_C \frac{1}{x}\,\mathrm{d}P_n(x)$. Therefore we can take $f_n(x) = f(x) = \frac{1}{x}$, and since the estimand is deterministic, we have a strong canonical estimation problem.

The second quantity of interest is the interval C, or equivalently its endpoints $\check{c}$ and $\hat{c}$. Note that, by construction, P is supported on C. Without loss of generality, we may assume that $\check{c}$ and $\hat{c}$ are respectively the essential infimum and essential supremum of $Z \sim P$. Therefore, note that $\left(\int x^{\pm q}\,\mathrm{d}P(x)\right)^{1/q}$ converges to the essential infimum $(-)$ or supremum $(+)$ as $q \to \infty$. We can therefore consider, for fixed $q \geq 1$, the strong canonical problems that ensue from the choices $f_n(x) = f(x) = x^{-q}$ and $f_n(x) = f(x) = x^q$. These, by themselves, are not sufficient to provide estimates for $\check{c}$ and $\hat{c}$. However if, in addition to consistency, we establish the convergence rates of their estimators, then we can apply our framework to estimate C, as we show in Section VI.

III. Solution Methodology

Our task now is to exhibit consistent estimators to canonical problems. We present here our abstract methodology, which we demonstrate concretely in Section V. The core of our approach consists of using the samples $X_{n,1}, \cdots, X_{n,n}$ to construct a random measure $\tilde{P}_n$ over $\mathbb{R}^+$, such that for almost every sample sequence, the sequence of measures $\{\tilde{P}_n\}$ converges weakly to P. We write: as $n \to \infty$

$$\tilde{P}_n \Rightarrow_{\text{a.s.}} P. \tag{5}$$

If we accomplish this, we can immediately suggest a consistent estimator under certain conditions, as expressed by Lemma 2. We will be interested in integrating functions against the measure $\tilde{P}_n$. However, since the support C of P is unknown, we first introduce the notion of a *tapered function* as a convenient way to control the region of integration. Given a real-valued function $g(x)$ on $\mathbb{R}^+$, for every $D \geq 1$ define its D-tapered version as:

$$g_D(x) \equiv \begin{cases} g(D^{-1}) & x < D^{-1} \\ g(x) & x \in [D^{-1}, D] \\ g(D) & x > D \end{cases}$$

If g is continuous on $(0, +\infty)$, then we can think of $g_D(x)$ as a bounded continuous extension of the restriction of g on $[D^{-1}, D]$ to all of $\mathbb{R}^+$.

Lemma 2. *Consider a canonical problem characterized by some f. Let the support C of a rare-events source be known up to an interval $[D^{-1}, D] \supseteq C$ for some $D > 1$. Then, if $\tilde{P}_n \Rightarrow_{\text{a.s.}} P$ as $n \to \infty$, we have that*

$$\hat{Y}_n = \int_{\mathbb{R}^+} f_D(x)\, \mathrm{d}\tilde{P}_n(x) \tag{6}$$

is a consistent estimator.

Furthermore, if f is bounded everywhere, we can make the uninformative choice $D = \infty$.

Proof: Since the tapered function f_D is continuous and bounded on $\mathbb{R}^+$, the almost sure weak convergence of $\tilde{P}_n$ to P implies that $\int_{\mathbb{R}^+} f_D \,\mathrm{d}\tilde{P}_n \to_{\text{a.s.}} \int_{\mathbb{R}^+} f_D \,\mathrm{d}P$. But since P is supported on C and f_D agrees with f on C, we have $\int_{\mathbb{R}^+} f_D \,\mathrm{d}P = \int_C f_D \,\mathrm{d}P = \int_C f \,\mathrm{d}P$. ∎

In general, however, we will be interested in problems where we do not have an *a priori* knowledge about the endpoints of C, and where an uninformative choice cannot be made because f is not bounded on $\mathbb{R}^+$, such as $f(x) = \log x$, $1/x$, or x^q. For these problems, we can apply our methodology of integrating against $\tilde{P}_n$ by first establishing a rate for the convergence of equation (5). We characterize such a rate using a sequence $K_n \to \infty$, such that:

$$K_n d_{\text{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0, \tag{7}$$

where d_{W} denotes the Wasserstein distance, which can be expressed in its dual forms:

$$\begin{aligned} d_{\text{W}}(\tilde{P}_n, P) &\equiv \int_{\mathbb{R}^+} |F(x; P_n) - F(x; P)| dx \\ &= \sup_{h \in \text{Lipschitz}(1)} \left| \int_{\mathbb{R}^+} h \,\mathrm{d}P_n - \int_{\mathbb{R}^+} h \,\mathrm{d}P \right|. \end{aligned} \tag{8}$$

In the remainder of the paper we will particularly focus on K_n of the form n^s for some $s > 0$.

In the following lemma, we describe how we can use convergence rates such as (7) to construct consistent estimators that work with no prior knowledge on C, for a large subclass of canonical problems.

Lemma 3. *Consider a canonical problem characterized by some f, which is Lipschitz on every closed interval $[a, b]$, $0 < a \leq b < \infty$. If $K_n d_{\text{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$ as $n \to \infty$, for some $K_n \to \infty$, then we can choose $D_n \to \infty$ such that*

$$\hat{Y}_n = \int_{\mathbb{R}^+} f_{D_n}(x)\, \mathrm{d}\tilde{P}_n(x) \tag{9}$$

is a consistent estimator. The growth of D_n controls the growth of the Lipschitz constant of f_{D_n}, which should be balanced with the convergence rate K_n. More precisely, $\hat{Y}_n$ in (9) is consitent for any $D_n \to \infty$ that additionally satisfies

$$\liminf_{n \to \infty} \frac{K_n}{\text{Lip}(f_{D_n})} > 0, \tag{10}$$

where $\text{Lip}(g)$ indicates the Lipschitz constant of g.

Proof: First note that for any $D \geq (\check{c}^{-1} \vee \hat{c})$, since P is supported on C and f_D agrees with f on C, we have:

$$\int_{\mathbb{R}^+} f_D \,\mathrm{d}P = \int_C f_D \,\mathrm{d}P = \int_C f \,\mathrm{d}P. \tag{11}$$

Then, using the fact that for every D, $f_D/\text{Lip}(f_D)$ is $\text{Lipschitz}(1)$, we can invoke the dual representation (8) of the Wasserstein distance to write:

$$K_n \sup_D \frac{1}{\text{Lip}(f_D)} \left| \int_{\mathbb{R}^+} f_D \,\mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} f_D \,\mathrm{d}P \right| \to_{\text{a.s.}} 0. \tag{12}$$

By combining equations (11) and (12), it follows that for any sequence $D_n \to \infty$, we have:

$$\frac{K_n}{\text{Lip}(f_{D_n})} \left| \int_{\mathbb{R}^+} f_{D_n} \,\mathrm{d}\tilde{P}_n - \int_C f \,\mathrm{d}P \right| \to_{\text{a.s.}} 0. \tag{13}$$

If furthermore D_n is chosen such that equation (10) is satisfied, then the factor $\frac{K_n}{\text{Lip}(f_{D_n})}$ is eventually bounded away from zero, and can be eliminated from equation (13) to lead to the convergence of the estimator. ∎

Of course, there may be more than one way in which one could construct $\tilde{P}_n$. In this paper, we focus on demonstrating the validity and usefulness of the methodology by providing two possible constructions. The results would remain valid regardless to the specific construction, and other constructions boasting more appealing properties, such as rates of convergence under more lenient assumptions, are welcome future contributions to this framework.

IV. The Good-Turing Pseudo-Empirical Measure

A. Definitions and Properties

The platform on which we build our estimation scheme is the Good-Turing estimator, and in particular its strong consistency established by WVK. In this section, we review the main definition and properties relevant to the rest of

the development. Let $B_{n,k}$ be the subset of symbols of A_n that appear exactly k times in the samples $X_{n,1}, \cdots, X_{n,n}$. The Good-Turing estimation problem, in reference to the pioneering work of Good in [4], is the estimation of the quantities $\gamma_{n,k} = p_n(B_{n,k})$, for each $k = 0, 1, \cdots, n$, that is the total probability of all symbols that appear exactly k times. We can group these with the notation $\gamma_n \equiv \{\gamma_{n,k}\}_{k\in\mathbb{N}_0}$, which we pad with zeros for $k > n$. In particular, Good suggests the following estimator.

Definition 5. Let $\varphi_{n,k} = |B_{n,k}|$ be the number of symbols of A_n that appear k times in $X_{n,1}, \cdots, X_{n,n}$. The *Good-Turing estimator* $\phi_n \equiv \{\phi_{n,k}\}_{k\in\mathbb{N}_0}$ of γ_n, for each $k \in \mathbb{N}_0$, is

$$\phi_{n,k} = \frac{(k+1)\varphi_{n,k+1}}{n}. \tag{14}$$

WVK establish a host of convergence properties for the Good-Turing estimation problem and the Good-Turing estimator. We group these in the following theorem.

Theorem 1. *Define the Poisson P-mixture $\lambda \equiv \{\lambda_k\}_{k\in\mathbb{N}_0}$ as, for each $k \in \mathbb{N}_0$:*

$$\lambda_k = \int_C \frac{x^k e^{-x}}{k!}\, \mathrm{d}P(x). \tag{15}$$

We then have the following results that determine the limiting behavior of γ_n, and the strong consistency of the Good-Turing estimator ϕ_n :

(i) *We have that $\gamma_{n,k} \to_{\text{a.s.}} \lambda_k$ and $\phi_{n,k} \to_{\text{a.s.}} \lambda_k$, and therefore $|\phi_{n,k} - \gamma_{n,k}| \to_{\text{a.s.}} 0$, pointwise for each $k \in \mathbb{N}_0$ as $n \to \infty$.*

(ii) *By Scheffé's theorem ([2], p. 215), it also follows that these convergences hold in L_1 almost surely, in that $\|\gamma_n - \lambda\|_1 \to_{\text{a.s.}} 0$ and $\|\phi_n - \lambda\|_1 \to_{\text{a.s.}} 0$, and therefore $\|\phi_n - \gamma_n\|_1 \to_{\text{a.s.}} 0$, as $n \to \infty$.*

B. Empirical Measure Analogy

The analogy that we would like to make in this section is the following. Assuming λ is given, one could take n i.i.d. samples from it, and form the empirical measure or the type, call it $\hat{\lambda}_n \equiv \{\hat{\lambda}_{n,k}\}_{k\in\mathbb{N}_0}$. Such an empirical measure would satisfy well-known statistical properties, in particular the strong law of large numbers would apply, and we would have $\hat{\lambda}_{n,k} \to_{\text{a.s.}} \lambda_k$. By Scheffé's theorem, L_1 convergence would also follow. It is evident from Theorem 1 that despite the fact that we do not have such a true empirical measure, the Good-Turing estimator ϕ_n behaves as one, and we may be justified to call it a *pseudo-empirical measure*.

Now observe that since, for discrete distributions, the total variation distance is related to the L_1 distance by $\sup_{B\subset\mathbb{N}_0} |\hat{\lambda}_n(B) - \lambda(B)| = \frac{1}{2}\|\hat{\lambda}_n - \lambda\|_1$, the true empirical measure also converges in total variation. As a special case, the Glivenko-Cantelli theorem applies in that $\sup_k |F(k;\lambda) - F(k;\hat{\lambda}_n)| \to_{\text{a.s.}} 0$. Recall that $F(.;.)$ denotes the cumulative of the second argument (a measure) evaluated at the first argument. In light of the above, this remains valid for the pseudo-empirical measure. However, for the classical empirical measure, we also have the *rate* of convergence in the Glivenko-Cantelli theorem, in the form of the Kolmogorov-Smirnov theorem and its variants for discrete distributions, see for example [5]. Such results are often formulated in terms of a convergence in probability of rate $\frac{1}{\sqrt{n}}$. So we next ask whether such rates hold for the pseudo-empirical measure as well.

We first note that the rare-events source model is lenient, in the sense that it does not impose any convergence rate on $P_n \Rightarrow P$. Therefore, convergence results that aim to parallel those of a true empirical measure will depend on assumptions on the rate of this core convergence. In particular, let us assume that we know something about the weak convergence rate of P_n to P in terms of the Wasserstein distance, in that we assume there exists an $r > 0$ such that

$$n^r d_{\text{W}}(P_n, P) \to 0.$$

For example, in Lemma 5, we will show that this holds true for a class of rare-events sources suggested by WVK.

Next, note that Lemma 11 in WVK gives the following useful concentration rate for the pseudo-empirical measure around its mean.

Lemma 4. *For any $\delta > 0$, $n^{1/2-\delta}\|\phi_n - \mathbf{E}[\phi_n]\|_1 \to_{\text{a.s.}} 0$.*

In the following statement, we show that a Kolmogorov-Smirnov-type convergence to λ does hold for the pseudo-empirical measure ϕ_n, with a rate that is essentially the slower of that of the concentration of Lemma 4 and that of the rare-events source itself.

Theorem 2. *Let $r > 0$ be such that $n^r d_{\text{W}}(P_n, P) \to 0$. Then for any $\delta > 0$, we have:*

$$n^{\min\{r,\, 1/2\}-\delta} \sup_k |F(k;\lambda) - F(k;\phi_n)| \to_{\text{a.s.}} 0. \tag{16}$$

Proof: For convenience, define $B_k \equiv \{0, \cdots, k\}$. The proof requires three approximations. The first is to approximate ϕ_n with $\mathbf{E}[\phi_n]$. This is already achieved using Lemma 4. Since the L_1 distance is twice the total variation distance, and specializing to the subsets B_k, we have that for all $\delta > 0$:

$$n^{1/2-\delta} \sup_k |F(k;\mathbf{E}[\phi_n]) - F(k;\phi_n)| \to_{\text{a.s.}} 0. \tag{17}$$

The next two approximations are *(a)* to approximate $\mathbf{E}[\phi_n]$ with a Poisson P_n-mixture (using the theory of Poisson approximation), and *(b)* to approximate the latter with λ, which is a Poisson P-mixture (using the convergence in $d_{\text{W}}(P_n, P)$).

Part (a) – For convenience, let π_n be a $\text{Poisson}(x)$ P_n-mixture, and let η_n be a $\text{Binomial}\left(\frac{x}{n}, n\right)$ P_n-mixture. One can show, as in the proof of Lemma 7 of WVK, that $\mathbf{E}[\phi_n]$ is a $\text{Binomial}\left(\frac{x}{n}, n-1\right)$ P_n-mixture. We first relate $\mathbf{E}[\phi_n]$ to η_n which is the natural candidate for Poisson approximation. We then use Le Cam's theorem to relate η_n to π_n.

We start with a general observation. Let $\mathscr{F} = \{f(\cdot;x) : x \in C\}$ and $\mathscr{G} = \{g(\cdot;x) : x \in C\}$ be two parametric classes of probability mass functions over $\mathbb{N}_0$, e.g. Poisson and Binomial, and let Q be a mixing distribution supported on C. Say that for some subset $B \subset \mathbb{N}_0$, we have the pointwise

bound $|f(B;x) - g(B;x)| \leq \ell(x)$. It follows that the mixture of the bound is also a bound on the mixture. More precisely:

$$\left| \int_C f(B;x) dQ(x) - \int_C g(B;x) dQ(x) \right| \leq \int_C \ell(x) dQ(x). \tag{18}$$

Note that if the pointwise bound above holds uniformly over B, then the same is true for the mixture bound. We will use this particularly with the subsets B_k, to bound the difference of cumulative distribution functions.

Now let $g_n(k;x)$ be the c.d.f. of a $\mathrm{Binomial}\left(\frac{x}{n}, n\right)$ random variable, and let $\tilde{g}_n(k;x)$ be the c.d.f. of a $\mathrm{Binomial}\left(\frac{x}{n}, n-1\right)$ random variable. For any given k, we have the following:

$$\begin{aligned}
&\left(1 - \frac{x}{n}\right) \tilde{g}_n(k;x) \\
&= \sum_{m=0}^{k} \frac{n-m}{n} \binom{n}{m} \left(\frac{x}{n}\right)^m \left(1 - \frac{x}{n}\right)^{n-m} \\
&= g_n(k;x) - \frac{1}{n} \sum_{m=0}^{k} m \binom{n}{m} \left(\frac{x}{n}\right)^m \left(1 - \frac{x}{n}\right)^{n-m}.
\end{aligned}$$

Using the facts that the sum is no larger than the mean and that $\tilde{g}_n(k;x) \leq 1$, it follows that for any given k we have:

$$\begin{aligned}
&|g_n(k;x) - \tilde{g}_n(k;x)| \\
&= \left| \frac{1}{n} \sum_{m=0}^{k} m \binom{n}{m} \left(\frac{x}{n}\right)^m \left(1 - \frac{x}{n}\right)^{n-m} - \frac{x}{n} \tilde{g}_n(k;x) \right| \\
&\leq \frac{x}{n}
\end{aligned}$$

Note that $\int_C g_n(k;x)\, \mathrm{d}P_n = F(k;\eta_n)$, the c.d.f. of η_n, and $\int_C \tilde{g}_n(k;x)\, \mathrm{d}P_n = F(k;\mathbf{E}[\phi_n])$, the c.d.f. of $\mathbf{E}[\phi_n]$. Using the observation leading to equation (18), it follows that:

$$\sup_k |F(k;\mathbf{E}[\phi_n]) - F(k;\eta_n)| \leq \frac{1}{n} \int_C x\, \mathrm{d}P_n(x) \leq \frac{\hat{c}}{n}. \tag{19}$$

Using Le Cam's theorem (see, for example, [6]), we know that the total variation distance, and hence the difference of probabilities assigned to any subset $B \subset \mathbb{N}_0$ by a $\mathrm{Poisson}(x)$ distribution and a $\mathrm{Binomial}\left(\frac{x}{n}, n\right)$ distribution is upper-bounded by $\frac{x^2}{n}$. We apply this to the subsets B_k, and use the observation leading to equation (18) once again to extend this result to the respective P_n-mixtures:

$$\sup_k |F(k;\pi_n) - F(k;\eta_n)| \leq \frac{1}{n} \int_C x^2\, \mathrm{d}P_n(x) \leq \frac{\hat{c}^2}{n}. \tag{20}$$

By combining equations (19) and (20), we deduce that for all $\delta > 0$:

$$n^{1-\delta} \sup_k |F(k;\mathbf{E}[\phi_n]) - F(k;\pi_n)| \to 0. \tag{21}$$

Part (b) – Now let $h(k;x)$ be the c.d.f. of a $\mathrm{Poisson}(x)$ random variable. Observe that:

$$\begin{aligned}
0 &\leq \frac{\mathrm{d}}{\mathrm{d}x} h(k;x) = \sum_{m=0}^{k} -\frac{x^m e^{-x}}{m!} + m\frac{x^{m-1}e^{-x}}{m!} \\
&\leq \frac{1}{x} \sum_{m=0}^{k} m \frac{x^m e^{-x}}{m!} = \frac{1}{x} \mathbf{E}\left[\mathrm{Poisson}(x)\right] = 1.
\end{aligned}$$

Therefore, when viewed as a function of x, $h(k;x)$ is a $\mathrm{Lipschitz}(1)$ function on C for all k. Using the dual representation of the Wasserstein distance, we then have:

$$\begin{aligned}
&\sup_k |F(k;\pi_n) - F(k;\lambda)| \\
&= \sup_k \left| \int_C h(k;x)\, \mathrm{d}P_n(x) - \int_C h(k;x)\, \mathrm{d}P(x) \right| \\
&\leq \sup_{h \in \mathrm{Lipschitz}(1)} \left| \int_C h\, \mathrm{d}P_n - \int_C h\, \mathrm{d}P \right| = d_W(P_n, P).
\end{aligned}$$

Using the assumption of the convergence rate of P_n to P, it follows that for all $\delta > 0$ we have:

$$n^{r-\delta} \sup_k |F(k;\pi_n) - F(k;\lambda)| \to 0. \tag{22}$$

The statement of the theorem follows by combining equations (17), (21), and (22). ∎

In a practical situation, one would expect that the rare-events source is well-behaved enough that $r > 1/2$, and that the bottleneck of Theorem 2 is given by the $1/2$ rate, and therefore we have a behavior that more closely parallels a true empirical measure. Indeed, some natural constructions obey this principle. Most trivially, for a sequence of uniform sources, e.g. if $p_n(a) = 1/n$, we have $P_n = P$, and therefore $r = \infty$. More generally, consider the following class of rare-events sources suggested by WVK.

Definition 6. Let g be a density on $[0,1]$ that is continuous Lebesgue almost everywhere, and such that $\check{c} \leq g(w) \leq \hat{c}$ for all $w \in [0,1]$. Let $A_n = \{1, \cdots, \lfloor \alpha n \rfloor\}$ for some $\alpha > 0$, and for every $a \in A_n$ let $p_n(a) = \int_{(a-1)/\lfloor \alpha n \rfloor}^{a/\lfloor \alpha n \rfloor} g(w)\, \mathrm{d}w$. One can then verify that $\{(A_n, p_n)\}$ is indeed a rare-events source, with P being the law of $g(W)$, where $W \sim g$. We call such a construction a *rare-events source obtained by quantizing* g.

Lemma 5. *Let g be a density as in Definition 6, and let $\{(A_n, p_n)\}$ be a rare-events source obtained by quantizing g. If g has finitely many discontinuities, and is Lipschitz within each interval of continuity, then for all $r < 1$:*

$$n^r d_{\mathrm{W}}(P_n, P) \to 0$$

Proof: Without loss of generality, assume $\alpha = 1$, and that the largest Lipschitz constant is 1. Consider the quantized density on $[0,1]$:

$$g_n(w) = n \int_{(\lceil wn \rceil - 1)/n}^{\lceil wn \rceil / n} g(v)\, \mathrm{d}v,$$

where the integral is against the Lebesgue measure. Then it follows that P_n is the law of $g_n(W_n)$, where $W_n \sim g_n$.

Say g has L discontinuities, and let D_n be the union of the L intervals of the form $[(a-1)/n, a/n]$ which contain these discontinuities. In all other intervals, we have that $|g(w) - g_n(w)| \leq 1/n$, using Lipschitz continuity and the intermediate value theorem. It follows that

$$\begin{aligned}
&\int_{[0,1]} |g(w) - g_n(w)|\, \mathrm{d}w \\
&= \int_{D_n} |g - g_n|\, \mathrm{d}w + \int_{[0,1] \setminus D_n} |g - g_n|\, \mathrm{d}w \leq \frac{L}{n} + \frac{1}{n}.
\end{aligned}$$

For any particular $x \in C$, let $B_x = \{w \in [0,1] : g(w) < x\}$. We then have

$$\begin{aligned} |F(x;P_n) - F(x;P)| &= \left| \int_{B_x} g(w) - g_n(w)\,\mathrm{d}w \right| \\ &\leq \int_{B_x} |g(w) - g_n(w)|\,\mathrm{d}w \leq \frac{L+1}{n}. \end{aligned}$$

By integrating over all x:

$$d_{\mathrm{W}}(P_n, P) = \int_C |F(x;P_n) - F(x;P)|\,\mathrm{d}x \leq \frac{(L+1)(\hat{c}-\check{c})}{n}.$$

Therefore the lemma follows. ∎

We end by remarking that the rare-events sources covered by Lemma 5 are rather general in nature. For example, all of the illustrative and numerical examples offered by WVK are special cases (more precisely, they have piecewise-constant g).

V. Constructing $\tilde{P}_n$ via Mixing Density Estimation

We would now like to address the task of using $X_{n,1}, \cdots, X_{n,n}$ to construct a sequence of probability measures $\tilde{P}_n$ that, for almost every sample sequence, converges weakly to P, as outlined in Section III. Since we have established the Good-Turing estimator as a pseudo-empirical measure issued from a Poisson P-mixture, in both consistency and rate, this is analogous to a mixture density estimation problem, with the true empirical measure replaced with the Good-Turing estimator ϕ_n.

We start by noting that the task is reasonable, because the mixing distribution in a Poisson mixture is identifiable from the mixture itself. This observation can be traced back to [7] and [8]. Then, the first natural approach is to use nonparametric maximum likelihood estimation. In Section V-A, we use Simar's work in [9] to construct a valid estimator in this framework. Unfortunately, to the best of the authors' knowledge, the maximum likelihood estimator does not have a well-studied rate of convergence on the recovered mixing distribution. In Section V-B we consider instead a minimum distance estimator, with which Chen gives optimal rates of convergence in [10], albeit by assuming finite support for P.

A. Maximum Likelihood Estimator

We first define the maximum likelihood estimator in our setting. Despite the fact that it is not, strictly speaking, maximizing a true likelihood, we keep this terminology in light of the origin of the construction.

Definition 7. Given the pseudo-empirical measure (Good-Turing estimator) ϕ_n the *maximum likelihood estimator* of the mixing distribution is a probability measure $\tilde{P}_n^{\mathrm{ML}}$ on $\mathbb{R}^+$ which maximizes the pseudo-likelihood as follows:

$$\tilde{P}_n^{\mathrm{ML}} \in \operatorname*{argmax}_Q \sum_{k=0}^{\infty} \phi_{n,k} \log \left(\int_0^\infty \frac{x^k e^{-x}}{k!}\,\mathrm{d}Q(x) \right). \quad (23)$$

It is not immediately clear whether $\tilde{P}_n^{\mathrm{ML}}$ exists or is unique. These questions were answered in the affirmative in [9]. On close examination, it is clear that these properties do not depend on whether we are using a pseudo-empirical measure instead of a true empirical measure. Hence they remain valid in our context. Next, we establish the main consistency statement.

Theorem 3. *For almost every sample sequence, the sequence $\{\tilde{P}_n^{ML}\}$ converges weakly to P as $n \to \infty$. We write this as $\tilde{P}_n^{ML} \Rightarrow_{\mathrm{a.s.}} P$.*

Proof: The main burden of proof is addressed by Theorem 1 in establishing the strong law of large numbers for the pseudo-empirical measure, and which is originally given in WVK's Proposition 7. Indeed, in Simar's proof ([9], Section 3.3, pp. 1203–1204), we only use the fact that $\phi_{n,k} \to_{\mathrm{a.s.}} \lambda_k$ for every $k \in \mathbb{N}_0$. The rest of the proof carries over, and the current theorem follows. ∎

It is worth noting that the consistency of the maximum likelihood estimator does not even require that condition (i) in the Definition 1 of the rare-events source to hold, since Theorem 1 in fact holds without that condition. In that sense, it is very general. However, when every neighborhood of 0 or ∞ has positive probability under P, it limits the types of functions that we can allow in the canonical problems, including sequence probabilities and entropies as discussed in WVK. When P is not compactly supported, it is also difficult to establish the rates of convergence.

B. Minimum Distance Estimator

We now define a minimum distance estimator for our setting. The reason that we suggest this alternate construction of $\tilde{P}_n$ is that it is useful to quantify the convergence rate to P, and the minimum distance estimator provides such a rate. However, it does so with the further assumption that P has a finite support, whose size is bounded by a known number m.

Also note that the definition of the estimator circumvents questions of existence by allowing for a margin of ϵ from the infimum, and does not necessarily call for uniqueness.

Definition 8. For a probability measure Q on $\mathbb{R}^+$, let $\pi(Q)$ denote the Poisson Q-mixture. Then, given the pseudo-empirical measure ϕ_n, a *minimum distance estimator* with precision ϵ is any probability measure $\tilde{P}_n^{\mathrm{MD},m,\epsilon}$ on $\mathbb{R}^+$ that satisfies

$$\begin{aligned} &\sup_k \left| F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon})) - F(k;\phi_n) \right| \\ &\qquad \leq \inf_Q \sup_k |F(k;\pi(Q)) - F(k;\phi_n)| + \epsilon, \end{aligned}$$

where the infimum is taken on probability measures supported on at most m points, on $\mathbb{R}^+$.

We now provide the main consistency and rate results associated with such estimators.

Theorem 4. *Let $r > 0$ be such that $n^r d_{\mathrm{W}}(P_n, P) \to 0$, and assume that it is known that P is supported on at most m points. Let $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}$ be a sequence of minimum distance estimators chosen such that $\epsilon_n < n^{-\min\{r,1/2\}}$. Then as $n \to \infty$, we have that for any $\delta > 0$:*

$$n^{\min\{r/2,1/4\}-\delta} d_{\mathrm{W}} \left(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}, P \right) \to_{\mathrm{a.s.}} 0. \quad (24)$$

Remark: Since d_{W} induces the weak convergence topology, it also follows that $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n} \Rightarrow_{\mathrm{a.s.}} P$.

Proof: To derive rate results in [10], Chen establishes a bound on the Wasserstein distance between mixing distributions, using the Kolmogorov-Smirnov distance between the c.d.f.s of the resulting mixtures. For this, he first introduces a notion of strong identifiability (Definition 2, p. 225), and shows that Poisson mixtures satisfy it (Section 4, p. 228). He then shows (in Lemma 2, p. 225) that if we have strongly identifiable mixtures and if two mixing distributions have a support of at most m points within a fixed compact set, such as C, then we can find a constant M (which depends non-constructively on m and C), such that for any two such mixing distributions Q_1 and Q_2, we have:

$$d_{\mathrm{W}}\left(Q_1, Q_2\right)^2 \leq M \sup_k |F(k; \pi(Q_1)) - F(k; \pi(Q_2))| \quad (25)$$

The main burden of proof therefore falls on our Theorem 2 in establishing a Kolmogorov-Smirnov-type convergence for the pseudo-empirical measure. The argument we present next is based on Chen's proof (Theorem 2, p. 226). We have:

$$\begin{aligned}
&\sup_k \left|F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k; \phi_n)\right| \\
&\quad \leq \sup_k \left|F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k; \lambda)\right| \\
&\qquad + \sup_k |F(k; \lambda) - F(k; \phi_n)| \\
&\quad \leq 2 \sup_k |F(k; \lambda) - F(k; \phi_n)| + \epsilon_n,
\end{aligned}$$

where the final inequality is due to the definition of $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}$. By Theorem 2, and by our choice of ϵ_n, it follows that for all $\delta > 0$, we have:

$$n^{\min\{r,1/2\}-2\delta} \sup_k \left|F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k; \phi_n)\right| \rightarrow_{\mathrm{a.s.}} 0. \quad (26)$$

By combining (25) and (26), the theorem follows . ■

Note that Chen's result can be used to show more. In particular, if we think of the true mixing distribution as residing in some neighborhood of a fixed distribution, then the convergence holds uniformly over that neighborhood. This may be interpreted as a form of robustness, but we do not dwell on it here.

VI. Applications

To solve canonical problems in the setting of Lemma 2, when an a priori bound on C is known or when f is bounded on $\mathbb{R}^+$, it suffices to construct a sequence of probability measures $\tilde{P}_n$ that weakly converges to P for almost every sample sequence. Since Theorem 3 provides such a sequence, we need not go further than that.

However, to work within the more general setting of Lemma 3, where no knowledge of C is assumed and f can be any locally Lipschitz function, we can use the result of Theorem 4. In this section, we start by illustrating this for some of the quantities considered by WVK. We then suggest two new applications: alphabet size and support interval estimation. We conclude by remarking on some algorithmic considerations.

A. Estimating Entropies and Probabilities

First consider the entropy of the source $H(p_n)$, and the associated problem, in normalized form, of estimating $Y_n^H \equiv H(p_n) - \log n$. One can then write:

$$Y_n^H = -\int_C \log x \, \mathrm{d}P_n(x),$$

and therefore, by comparing to equation (1) with $f_n(x) = f(x) = -log(x)$, we have a canonical estimation problem, and since Y_n^H is deterministic, it is also strong. If we have a bound on C, we can use Lemma 2. Otherwise, note that on intervals of the form $[D^{-1}, D]$, $\log x$ is D-Lipshitz. Therefore if for some $s > 0$, $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \rightarrow_{\mathrm{a.s.}} 0$, as given by Theorem 4 for example, then we can apply Lemma 3 using $D_n = n^s$. If s exists but is unknown, we can still apply Lemma 3 using any sequence that is $o(n^s)$, such as $D_n = e^{\log^\epsilon n}$, for some $\epsilon > 0$. The consistent estimator becomes:

$$\hat{Y}_n^H \equiv -\int_{\mathbb{R}^+} \log_{D_n} x \, \mathrm{d}\tilde{P}_n(x). \quad (27)$$

Next consider the probability of the sequence $p_n(X_{n,1}, \cdots, X_{n,n})$, and the associated normalized problem of estimating $Y_n^p \equiv \frac{1}{n} \log p_n(X_{n,1}, \cdots, X_{n,n}) + \log n$. We have (WVK, Lemma 5):

$$\begin{aligned}
\mathbf{E}[Y_n^p] &= \mathbf{E}[\log p_n(X_n)] + \log n \\
&= \int_C \log x \, \mathrm{d}P_n(x),
\end{aligned}$$

and therefore we also have a canonical estimation problem. Using McDiarmid's theorem, one can also show that (WVK, Lemma 6) $|\mathbf{E}[Y_n^p] - Y_n^p| \rightarrow_{\mathrm{a.s.}} 0$, and therefore we once again have a strong canonical estimation problem, and we can construct a consistent estimator as in the case of entropy. Referring to equation (27), we have $\hat{Y}_n^p \equiv -\hat{Y}_n^H$.

B. Estimating the Alphabet Size

Consider the size of the alphabet $|A_n|$. Since the model describes large, asymptotically infinite, alphabets, we look at the normalized problem of estimating $Y_n^A = |A_n|/n$. We have (cf. [3]):

$$\begin{aligned}
Y_n^A &= \frac{1}{n} \sum_{a \in A} 1 = \sum_{a \in A} \frac{p_n(a)}{n p_n(a)} \\
&= \int_C \frac{1}{x} \, \mathrm{d}P_n(x).
\end{aligned}$$

Once again, having a deterministic sequence of the form of (1) with $f_n(x) = f(x) = 1/x$, it follows that $\{Y_n^A\}_{n \in \mathbb{N}}$ is a strong canonical problem. If we have a bound on C, we can use Lemma 2. Otherwise, note that on intervals of the form $[D^{-1}, D]$, $1/x$ is D^2-Lipshitz. Therefore if for some $s > 0$, $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \rightarrow_{\mathrm{a.s.}} 0$, as given by Theorem 4 for example, then we can apply Lemma 3 using $D_n = n^{s/2}$. As in Section VI-A, if s exists but is unknown, we can still apply Lemma 3 using any sequence that is $o(n^s)$, such as $D_n = e^{\log^\epsilon n}$, for some $\epsilon > 0$. The consistent estimator becomes:

$$\hat{Y}_n^A \equiv \int_{\mathbb{R}^+} x_{D_n}^{-1} \, \mathrm{d}\tilde{P}_n(x). \quad (28)$$

C. Estimating the Support Interval

As discussed in Section II-B, estimating the support interval is not a canonical problem per se. However, we show here that we can extend the framework in a straightforward fashion to provide consistent estimators of both $\check{c}$ and $\hat{c}$.

Lemma 6. *Let $\tilde{P}_n \Rightarrow_{\text{a.s.}} P$ such that for some $s > 0$, we have $n^s d_{\text{W}}(\tilde{P}_n, P) \rightarrow_{\text{a.s.}} 0$. This is particularly true under the conditions of Theorem 4. Given $q \neq 0$ and $D \geq 1$, let x_D^q denote the D-tapered version of x^q.*

If $q_n = \log n / \log\log n$ and $D_n = n^{s/(2q_n)}$, then we have: as $n \to \infty$,

$$\left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \, \mathrm{d}\tilde{P}_n(x)\right)^{1/q_n} \rightarrow_{\text{a.s.}} \check{c}$$

$$\text{and} \quad \left(\int_{\mathbb{R}^+} x_{D_n}^{q_n} \, \mathrm{d}\tilde{P}_n(x)\right)^{1/q_n} \rightarrow_{\text{a.s.}} \hat{c}.$$

Proof: For conciseness, let us drop the argument of the probability measures, and write $\mathrm{d}P$ for $\mathrm{d}P(x)$. We provide the proof only for $\check{c}$, since the argument is analogous for $\hat{c}$. Recall that $\check{c}$ is the essential infimum of a random variable $Z \sim P$. Therefore, for any $D \geq (\check{c}^{-1} \vee \hat{c})$, we have:

$$\left(\int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}P\right)^{1/q} \to \check{c} \qquad \text{as } q \to \infty. \tag{29}$$

In the absence of a rate of convergence, we cannot simply plug in $\tilde{P}_n$. But since we know that $n^s d_{\text{W}}(\tilde{P}_n, P) \rightarrow_{\text{a.s.}} 0$, we can use the dual representation of the Wasserstein distance and the fact that for every q and D the function $\frac{1}{q} D^{-1-q} x_D^{-q}$ is $\text{Lipschitz}(1)$ over $\mathbb{R}^+$ to state: as $n \to \infty$,

$$n^s \sup_{q,D} \frac{D^{-1-q}}{q} \left| \int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}P \right| \rightarrow_{\text{a.s.}} 0. \tag{30}$$

We now want to relate this to the difference of the q^{th} roots. Note that each of the integrals in (30) is bounded from below by D^{-q}. Using this and the fact that for any a and $b > 0$ we have $\left|a^{1/q} - b^{1/q}\right| \leq \frac{1}{q}(a \wedge b)^{\frac{1}{q}-1} |a - b|$, we can write:

$$\begin{aligned} &\left| \left(\int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}\tilde{P}_n\right)^{1/q} - \left(\int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}P\right)^{1/q} \right| \\ &\quad \leq \quad D^{2q} \cdot \frac{D^{-1-q}}{q} \left| \int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_D^{-q} \, \mathrm{d}P \right|. \end{aligned}$$

The choices $q_n = \log n / \log\log n$ and $D_n = n^{s/(2q_n)}$, allow us to have $D_n^{2q_n} = n^s$, and yet guarantee that as $n \to \infty$ both q_n and $D_n \to \infty$. With this, we can use the convergence of equation (30), to state: as $n \to \infty$,

$$\begin{aligned} &\left| \left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \, \mathrm{d}\tilde{P}_n\right)^{1/q_n} - \left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \, \mathrm{d}P\right)^{1/q_n} \right| \\ &\leq n^s \frac{D_n^{-1-q_n}}{q_n} \left| \int_{\mathbb{R}^+} x_{D_n}^{-q_n} \, \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_{D_n}^{-q_n} \, \mathrm{d}P \right| \rightarrow_{\text{a.s.}} 0. \end{aligned} \tag{31}$$

We then combine (29) and (31) to complete the proof. ∎

Remarks. Note the following:

(i) Other scaling schemes can be devised for q_n and D_n, as long as they both grow to ∞ as $n \to \infty$, yet $D_n^{2q_n}$ remains at most $\mathcal{O}(n^s)$.

(ii) If a bound $[D_{\min}, D_{\max}] \supset C$ is already known, then we can taper x^q accordingly, without growing D_n. In this case, we can also speed up the rate of convergence by choosing $q_n = \frac{s}{2} \log n / \log \frac{D_{\max}}{D_{\min}}$.

(iii) If only an upper bound or only a lower bound is known, we can taper x^q accordingly, and only grow/shrink the missing bound. In this case we leave $q_n = \log n / \log\log n$ as in the Lemma.

(iv) In the Lemma and the alternatives in these remarks, if s is unknown we can replace it wherever it appears (together with constant factors) with a suitably decaying term, that guarantees the behavior of remark (i). For example, in the Lemma, we can choose $D_n = n^{1/(q_n \sqrt{\log\log n})}$, since then $D_n^{2q_n}$ becomes $o(n^s)$ for any s, and the proof applies.

D. Algorithmic Considerations

One of the appealing properties of the maximum likelihood estimator is that, by a result of Simar in [9], it is supported on finitely many points. Simar also suggests a particular algorithm for obtaining the $\tilde{P}_n^{\text{MLE}}$, the convergence of which was later established in [11], with further improvements. One can also solve for the MLE using the EM algorithm, as reviewed in [12]. Penalized variants are also suggested, such as in [13]. The literature on the non-parametric maximum likelihood estimator for mixtures is indeed very rich. As for the minimum distance estimator, in [10] Chen suggests variants of the work in [14], where they use algorithms based on linear programming.

References

[1] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3207–3229, 2011.

[2] P. Billingsley, *Probability and Measure*, 3rd ed. NY: Wiley, 1995.

[3] S. Bhat and R. Sproat, "Knowing the unseen: estimating vocabulary size over unseen samples," in *Proceedings of the ACL*, Suntec, Singapore, Aug. 2009, pp. 109–117.

[4] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 16, pp. 237–264, 1953.

[5] C. L. Wood and M. M. Altavela, "Large-sample results for kolmogorov-smirnov statistics for discrete distributions," *Biometrika*, vol. 65, no. 1, pp. 235–239, 1978.

[6] M. J. Steele, "Le cam's inequality and poisson approximations," *American Mathematical Monthly*, vol. 101, no. 1, pp. 48–54, 1994.

[7] H. Teicher, "Identifiability of mixtures," *Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 244–248, 1961.

[8] W. Feller, "On a general class of "contagious" distributions," *Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 389–400, 1943.

[9] L. Simar, "Maximum likelihood estimation of a compound Poisson process," *Annals of Statistics*, vol. 4, no. 6, pp. 1200–1209, 1976.

[10] J. Chen, "Optimal rate of convergence for finite mixture models," *Annals of Statistics*, vol. 23, no. 1, pp. 221–233, 1995.

[11] D. Böhning, "Convergence of Simar's algorithm for finding the maximum likelihood estimator of a compound poisson process," *Annals of Statistics*, vol. 10, no. 3, pp. 1006–1008, 1982.

[12] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[13] B. G. Leroux, "Consistent estimation of a mixing distribution," *Annals of Statistics*, vol. 20, no. 3, pp. 1350–1360, 1992.

[14] J. J. Deely and R. L. Kruse, "Construction of sequences estimating the mixing distribution," *Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 286–288, 1968.