entropy

MDPI

Article

Shannon Entropy Estimation in ∞-Alphabets from Convergence Results: Studying Plug-In Estimators

Jorge F. Silva

Information and Decision System Group, Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, Santiago 7591538, Chile; josilva@ing.uchile.cl

Received: 12 April 2018; Accepted: 18 May 2018; Published: 23 May 2018

Abstract: This work addresses the problem of Shannon entropy estimation in countably infinite alphabets studying and adopting some recent convergence results of the entropy functional, which is known to be a discontinuous function in the space of probabilities in ∞-alphabets. Sufficient conditions for the convergence of the entropy are used in conjunction with some deviation inequalities (including scenarios with both finitely and infinitely supported assumptions on the target distribution). From this perspective, four plug-in histogram-based estimators are studied showing that convergence results are instrumental to derive new strong consistent estimators for the entropy. The main application of this methodology is a new data-driven partition (plug-in) estimator. This scheme uses the data to restrict the support where the distribution is estimated by finding an optimal balance between estimation and approximation errors. The proposed scheme offers a consistent (distribution-free) estimator of the entropy in ∞-alphabets and optimal rates of convergence under certain regularity conditions on the problem (finite and unknown supported assumptions and tail bounded conditions on the target distribution).

Keywords: Shannon entropy estimation; countably infinite alphabets; entropy convergence results; statistical learning; histogram-based estimators; data-driven partitions; strong consistency; rates of convergence

1. Introduction

Shannon entropy estimation has a long history in information theory, statistics, and computer science [1]. Entropy and related information measures (conditional entropy and mutual information) have a fundamental role in information theory and statistics [2,3] and, as a consequence, it has found numerous applications in learning and decision making tasks [4–15]. In many of these contexts, distributions are not available and the entropy needs to be estimated from empirical data. This problem belongs to the category of scalar functional estimation that has been thoroughly studied in non-parametric statistics.

Starting with the finite alphabet scenario, the classical plug-in estimator (i.e., the empirical distribution evaluated on the functional) is well known to be consistent, minimax optimal, and asymptotically efficient [16] (Sections 8.7–8.9). More recent research has focused on looking at the so-called large alphabet (or large dimensional) regime, meaning a non-asymptotic under-sampling regime where the number of samples n is on the order of, or even smaller than, the size of the alphabet denoted by k. In this context, it has been shown that the classical plug-in estimator is sub-optimal as it suffers from severe bias [17,18]. For characterizing optimality in this high dimensional context, a non-asymptotic minimax mean square error analysis (under a finite n and k) has been explored by several authors [17–21] considering the minimax risk

$$R^*(k,n) = \inf_{\hat{H}(\cdot)} \sup_{\mu \in \mathcal{P}(k)} \mathbb{E}_{X_1,\ldots X_n \sim \mu^n} \left\{ \left(\hat{H}(X_1,\ldots,X_n) - H(\mu) \right)^2 \right\}$$

where $\mathcal{P}(k)$ denotes the collection of probabilities on $[k] \equiv \{1, \ldots, k\}$ and $H(\mu)$ is the entropy of μ (details in Section 2). Paninski [19] first showed that it was possible to construct an entropy estimator that uses a sub-linear sampling size to achieve minimax consistency when k goes to infinity, in the sense that there is a sequence $(n_k) = o(k)$ where $R^*(k, n_k) \longrightarrow 0$ as k goes to infinity. A set of results by Valiant et al. [20,21] shows that the optimal scaling of the sampling size with respect to k is $O(k/\log(k))$, to achieve the aforementioned asymptotic consistency for entropy estimation. A refined set of results for the complete characterization of $R^*(k, n)$, the specific scaling of the sampling complexity, and the achievability of the obtained minimax L_2 risk for the family $\{\mathcal{P}(k) : k \geq 1\}$ with practical estimators have been presented in [17,18]. On the other hand, it is well-known that the problem of estimating the distribution (consistently in total variation) in finite alphabets requires a sampling complexity that scales as $O(k)$ [22]. Consequently, in finite alphabets the task of entropy estimation is simpler than estimating the distribution in terms of sampling complexity. These findings are consistent with the observation that the entropy is a continuous functional of the space of distributions (in the total variational distance sense) for the finite alphabet case [2,23–25].

1.1. The Challenging Infinite Alphabet Learning Scenario

In this work, we are interested in the countably infinite alphabet scenario, i.e., on the estimation of the entropy when the alphabet is countably infinite and we have a finite number of samples. This problem can be seen as an infinite dimensional regime as the size of the alphabet goes unbounded and n is kept finite for the analysis, which differs from the large dimensional regime mentioned above. As argued in [26] (Section IV), this is a challenging non-parametric learning problem because some of the finite alphabet properties of the entropy do not extend to this infinite dimensional context. Notably, it has been shown that the Shannon entropy is not a continuous functional with respect to the total variational distance in infinite alphabets [24,26,27]. In particular, Ho et al. [24] (Theorem 2) showed concrete examples where convergence in χ^2-divergence and in direct information divergence (I-divergence) of a set of distributions to a limit, both stronger than total variational convergence [23,28], do not imply the convergence of the entropy. In addition, Harremoës [27] showed the discontinuity of the entropy with respect to the reverse I-divergence [29], and consequently, with respect to the total variational distance (the distinction between reverse and direct I-divergence was pointed out in the work of Barron et al. [29]). In entropy estimation, the discontinuity of the entropy implies that the minimax mean square error goes unbounded, i.e.,

$$R_n^* = \inf_{\hat{H}(\cdot)} \sup_{\mu \in \mathcal{H}(\mathbb{X})} \mathbb{E}_{X_1, \ldots X_n \sim \mu^n} \left\{ \left(\hat{H}(X_1, \ldots, X_n) - H(\mu) \right)^2 \right\} = \infty,$$

where $\mathcal{H}(\mathbb{X})$ denotes the family of finite entropy distribution over the countable alphabet set $\mathbb{X}$ (the proof of this result follows from [26] (Theorem 1) and the argument is presented in Appendix A). Consequently, there is no universal minimax consistent estimator (in the mean square error sense) of the entropy over the family of finite entropy distributions.

Considering a sample-wise (or point-wise) convergence to zero of the estimation error (instead of the minimax expected error analysis mentioned above), Antos et al. [30] (Theorem 2 and Corollary 1) show the remarkable result that the classical plug-in estimate is strongly consistent and consistent in the mean square error sense for any finite entropy distribution (point-wise). Then, the classical plug-in entropy estimator is universal, meaning that the convergence to the right limiting value $H(\mu)$ is achieved almost surely despite the discontinuity of the entropy. Moving on the analysis of the (point-wise) rate of convergence of the estimation error, Antos et al. [30] (Theorem 3) present a finite length lower bound for the error of any arbitrary estimation scheme, showing as a corollary that no universal rate of convergence (to zero) can be achieved for entropy estimation in infinite alphabets [30] (Theorem 4). Finally, constraining the problem to a family of distributions with specific power tail bounded conditions, Antos et al. [30] (Theorem 7) present a finite length expression for the rate of convergence of the estimation error of the classical plug-in estimate.

1.2. From Convergence Results to Entropy Estimation

In view of the discontinuity of the entropy in ∞-alphabets [24] and the results that guarantee entropy convergence [25–27,31], this work revisits the problem of point-wise almost-sure entropy estimation in ∞-alphabets from the perspective of studying and applying entropy convergence results and their derived bounds [25,26,31]. Importantly, entropy convergence results have established concrete conditions on both the limiting distribution μ and the way a sequence of distributions $\{\mu_n : n \geq 0\}$ converges to μ such that $\lim_{n\to\infty} H(\mu_n) = H(\mu)$ is satisfied. The natural observation that motivates this work is that consistency is basically a convergence to the true entropy value that happens with probability one. Then our main conjecture is that putting these conditions in the context of a learning task, i.e., where $\{\mu_n : n \geq 0\}$ is a random sequence of distributions driven by the classical empirical process, will offer the possibility to study a broad family of plug-in estimators with the objective to derive new strong consistency and rates of convergence results. On the practical side, this work proposes and analizes a data-driven histogram-based estimator as a key learning scheme, since this approach offers the flexibility to adapt to learning task when appropriate bounds for the estimation and approximation errors are derived.

1.3. Contributions

We begin revisiting the classical plug-in entropy estimator considering the relevant scenario where μ (the unknown distribution that produces the i.i.d. samples) has a finite but arbitrary large and unknown support. This is declared to be a challenging problem by Ho and Yeung [26] (Theorem 13) because of the discontinuity of the entropy. Finite-length (non-asymptotic) deviation inequalities and intervals of confidence are derived extending the results presented in [26] (Section IV). From this, it is shown that the classical plug-in estimate achieves optimal rates of convergence. Relaxing the finite support restriction on μ, two concrete histogram-based plug-in estimators are presented, one built upon the celebrated Barron-Györfi-van der Meulen histogram-based approach [29,32,33]; and the other on a data-driven partition of the space [34–36]. For the Barron plug-in scheme, almost-sure consistency is shown for entropy estimation and distribution estimation in direct I-divergence under some mild support conditions on μ. For the data-driven partition scheme, the main context of application of this work, it is shown that this estimator is strongly consistent distribution-free, matching the universal result obtained for the classical plug-in approach in [30]. Furthermore, new almost-sure rates of convergence results (in the estimation error) are obtained for distributions with finite but unknown support and for families of distributions with power and exponential tail dominating conditions. In this context, our results show that this adaptive scheme has a concrete design solution that offers very good convergence rate of the overall estimation error, as it approaches the rate $O(1/\sqrt{n})$ that is considered optimal for the finite alphabet case [16]. Importantly, the parameter selection of this scheme relies on, first, obtaining expressions to bound the estimation and approximation errors and, second, finding the optimal balance between these two learning errors.

1.4. Organization

The rest of the paper is organized as follows. Section 2 introduces some basic concepts, notation, and summarizes the main entropy convergence results used in this work. Sections 3–5 state and elaborate the main results of this work. Discussion of the results and final remarks are given in Section 6. The technical derivation of the main results are presented in Section 7. Finally, proofs of auxiliary results are relegated to the Appendix Section.

2. Preliminaries

Let $\mathbb{X}$ be a countably infinite set and let $\mathcal{P}(\mathbb{X})$ denote the collection of probability measures in $\mathbb{X}$. For μ and v in $\mathcal{P}(\mathbb{X})$, and μ absolutely continuous with respect to v (i.e., $\mu \ll v$), $\frac{d\mu}{dv}(x)$ denotes the *Radon-Nikodym* (RN) derivative of μ with respect to v. Every $\mu \in \mathcal{P}(\mathbb{X})$ is equipped with its

probability mass function (pmf) that we denote by $f_\mu(x) \equiv \mu(\{x\}), \forall x \in \mathbb{X}$. Finally, for any $\mu \in \mathcal{P}(\mathbb{X})$, $A_\mu \equiv \{x \in \mathbb{X} : f_\mu(x) > 0\}$ denotes its support and

$$\mathcal{F}(\mathbb{X}) \equiv \{\mu \in \mathcal{P}(\mathbb{X}) : |A_\mu| < \infty\} \tag{1}$$

denotes the collection of probabilities with finite support.

Let μ and v be in $\mathcal{P}(\mathbb{X})$, then the total variation distance of μ and v is given by [28]

$$V(\mu, v) \equiv \sup_{A \in 2^{\mathbb{X}}} |v(A) - \mu(A)|, \tag{2}$$

where $2^{\mathbb{X}}$ denotes the subsets of $\mathbb{X}$. The Kullback–Leibler divergence or I-divergence of μ with respect to v is given by

$$D(\mu||v) \equiv \sum_{x \in A_\mu} f_\mu(x) \log \frac{f_\mu(x)}{f_v(x)} \geq 0, \tag{3}$$

when $\mu \ll v$, while $D(\mu||v)$ is set to infinite, otherwise [37].

The Shannon entropy of $\mu \in \mathcal{P}(\mathbb{X})$ is given by [1,2,38]:

$$H(\mu) \equiv -\sum_{x \in A_\mu} f_\mu(x) \log f_\mu(x) \geq 0. \tag{4}$$

In this context, let $\mathcal{H}(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ be the collection of probabilities where (4) is finite, let $\mathcal{AC}(\mathbb{X}|v) \subset \mathcal{P}(\mathbb{X})$ denote the collection of probabilities absolutely continuous with respect to $v \in \mathcal{P}(\mathbb{X})$, and let $\mathcal{H}(\mathbb{X}|v) \subset \mathcal{AC}(\mathbb{X}|v)$ denote the collection of probabilities where (3) is finite for $v \in \mathcal{P}(\mathbb{X})$.

Concerning convergence, a sequence $\{\mu_n : n \in \mathbb{N}\} \subset \mathcal{P}(\mathbb{X})$ is said to converge in total variation to $\mu \in \mathcal{P}(\mathbb{X})$ if

$$\lim_{n \to \infty} V(\mu_n, \mu) = 0. \tag{5}$$

For countable alphabets, ref. [31] (Lemma 3) shows that the convergence in total variation is equivalent to the weak convergence, which is denoted here by $\mu_n \Rightarrow \mu$, and the point-wise convergence of the pmf's. Furthermore, from (2), the convergence in total variation implies the uniform convergence of the pmf's, i.e, $\lim_{n\to\infty} \sup_{x \in \mathbb{X}} |\mu_n(\{x\}) - \mu(\{x\})| = 0$. Therefore, in this countable case, all the four previously mentioned notions of convergence are equivalent: total variation, weak convergence, point-wise convergence of the pmf's, and uniform convergence of the pmf's.

We conclude with the convergence in I-divergence introduced by Barron et al. [29]. It is said that $\{\mu_n : n \in \mathbb{N}\}$ converges to μ in direct and in reverse I-divergence if $\lim_{n\to\infty} D(\mu||\mu_n) = 0$ and $\lim_{n\to\infty} D(\mu_n||\mu) = 0$, respectively. From Pinsker's inequality [39–41], the convergence in I-divergence implies the weak convergence in (5), where it is known that the converse result is not true [27].

2.1. Convergence Results for the Shannon Entropy

The discontinuity of the entropy in ∞-alphabets raises the problem of finding conditions under which convergence of the entropy can be obtained. On this topic, Ho et al. [26] have studied the interplay between entropy and the total variation distance, specifying conditions for convergence by assuming a finite support on the involved distributions. On the other hand, Harremoës [27] (Theorem 21) obtained convergence of the entropy by imposing a power dominating condition [27] (Definition 17) on the limiting probability measure μ for all the sequences $\{\mu_n : n \geq 0\}$ converging in reverse I-divergence to μ [29]. More recently, Silva et al. [25] have addressed the entropy convergence studying a number of new settings that involve conditions on the limiting measure μ, as well as the way the sequence $\{\mu_n : n \geq 0\}$ converges to μ in the space of distributions. These results offer sufficient conditions where the entropy evaluated in a sequence of distributions converges to the

entropy of its limiting distribution and, consequently, the possibility of applying these when analyzing plug-in entropy estimators. The results used in this work are summarized in the rest of this section.

Let us begin with the case when $\mu \in \mathcal{F}(\mathbb{X})$, i.e., when the support of the limiting measure is finite and unknown.

Proposition 1. *Let us assume that* $\mu \in \mathcal{F}(\mathbb{X})$ *and* $\{\mu_n : n \in \mathbb{N}\} \subset \mathcal{AC}(\mathbb{X}|\mu)$. *If* $\mu_n \Rightarrow \mu$, *then* $\lim_{n\to\infty} D(\mu_n||\mu) = 0$ *and* $\lim_{n\to\infty} H(\mu_n) = H(\mu)$.

This result is well-known because when $A_{\mu_n} \subset A_\mu$ for all n, the scenario reduces to the finite alphabet case, where the entropy is known to be continuous [2,23]. Since we obtain two inequalities that are used in the following sections, a simple proof is provided here.

Proof. μ and μ_n belong to $\mathcal{H}(\mathbb{X})$ from the finite-supported assumption. The same argument can be used to show that $D(\mu_n||\mu) < \infty$, since $\mu_n \ll \mu$ for all n. Let us consider the following identity:

$$H(\mu) - H(\mu_n) = \sum_{x\in A_\mu} (f_{\mu_n}(x) - f_\mu(x)) \log f_\mu(x) + D(\mu_n||\mu). \tag{6}$$

The first term on the right hand side (RHS) of (6) is upper bounded by $\mathbf{M}_\mu \cdot V(\mu_n, \mu)$ where

$$\mathbf{M}_\mu = \log \frac{1}{\mathbf{m}_\mu} \equiv \sup_{x\in A_\mu} |\log \mu(\{x\})| < \infty. \tag{7}$$

For the second term, we have that

$$\begin{aligned} D(\mu_n||\mu) &\leq \log e \cdot \sum_{x\in A_{\mu_n}} f_{\mu_n}(x) \left| \frac{f_{\mu_n}(x)}{f_\mu(x)} - 1 \right| \\ &\leq \frac{\log e}{\mathbf{m}_\mu} \cdot \sup_{x\in A_\mu} \left| f_{\mu_n}(x) - f_\mu(x) \right| \leq \frac{\log e}{\mathbf{m}_\mu} \cdot V(\mu_n, \mu). \end{aligned} \tag{8}$$

and, consequently,

$$|H(\mu) - H(\mu_n)| \leq \left[\mathbf{M}_\mu + \frac{\log e}{\mathbf{m}_\mu} \right] \cdot V(\mu_n, \mu). \tag{9}$$

□

Under the assumptions of Proposition 1, we note that the reverse I-divergence and the entropy difference are bounded by the total variation by (8) and (9), respectively. Note, however, that these bounds are a distribution-dependent function of $\mathbf{m}_\mu(\mathbf{M}_\mu)$ in (7) (it is direct to show that $\mathbf{m}_\mu(\mathbf{M}_\mu) < \infty$ if, and only if, $\mu \in \mathcal{F}(\mathbb{X})$). The next result relaxes the assumption that $\mu_n \ll \mu$ and offers a necessary and sufficient condition for the convergence of the entropy.

Lemma 1. *Ref. [25] (Theorem 1) Let* $\mu \in \mathcal{F}(\mathbb{X})$ *and* $\{\mu_n : n \in \mathbb{N}\} \subset \mathcal{F}(\mathbb{X})$. *If* $\mu_n \Rightarrow \mu$, *then there exists* $N > 0$ *such that* $\mu \ll \mu_n \ \forall n \geq N$, *and*

$$\lim_{n\to\infty} D(\mu||\mu_n) = 0.$$

Furthermore, $\lim_{n\to\infty} H(\mu_n) = H(\mu)$, *if and only if,*

$$\lim_{n\to\infty} \mu_n \left(A_{\mu_n} \setminus A_\mu \right) \cdot H\left(\mu_n \left(\cdot | A_{\mu_n} \setminus A_\mu \right) \right) = 0 \Leftrightarrow \lim_{n\to\infty} \sum_{x\in A_{\mu_n}\setminus A_\mu} f_{\mu_n}(x) \log \frac{1}{f_{\mu_n}(x)} = 0, \tag{10}$$

where $\mu(\cdot|B)$ *denotes the conditional probability of* μ *given the event* $B \subset \mathbb{X}$.,

Lemma 1 tells us that to achieve entropy convergence (on top of the weak convergence), it is necessary and sufficient to ask for a vanishing expression (with n) of the entropy of μ_n restricted to the elements of the set $A_{\mu_n} \setminus A_\mu$. Two remarks about this result are: (1) The convergence in direct I-divergence does not imply the convergence of the entropy (concrete examples are presented in [24] (Section III) and [25]); (2) Under the assumption that $\mu \in \mathcal{F}(\mathbb{X})$, μ is eventually absolutely continuous with respect to μ_n, and the convergence in total variations is equivalent to the convergence in direct I-divergence.

This section concludes with the case when the support of μ is infinite and unknown, i.e., $\left|A_\mu\right| = \infty$. In this context, two results are highlighted:

Lemma 2. *Ref. [31] (Theorem 4) Let us consider that $\mu \in \mathcal{H}(\mathbb{X})$ and $\{\mu_n : n \geq 0\} \subset \mathcal{AC}(\mathbb{X}|\mu)$. If $\mu_n \Rightarrow \mu$ and*

$$M \equiv \sup_{n\geq 1} \sup_{x \in A_\mu} \frac{f_{\mu_n}(x)}{f_\mu(x)} < \infty, \tag{11}$$

then $\mu_n \in \mathcal{H}(\mathbb{X}) \cap \mathcal{H}(\mathbb{X}|\mu)$ for all n and it follows that

$$\lim_{n\to\infty} D(\mu_n||\mu) = 0 \text{ and } \lim_{n\to\infty} H(\mu_n) = H(\mu).$$

Interpreting Lemma 2 we have that, to obtain the convergence of the entropy functional (without imposing a finite support assumption on μ), a uniform bounding condition (UBC) μ-almost everywhere was added in (11). By adding this UBC, the convergence on reverse I-divergence is also obtained as a byproduct. Finally, when $\mu \ll \mu_n$ for all n, the following result is considered:

Lemma 3. *Ref. [25] (Theorem 3) Let $\mu \in \mathcal{H}(\mathbb{X})$ and a sequence of measures $\{\mu_n : n \geq 1\} \subset \mathcal{H}(\mathbb{X})$ such that $\mu \ll \mu_n$ for all $n \geq 1$. If $\mu_n \Rightarrow \mu$ and*

$$\sup_{n\geq 1} \sup_{x \in A_\mu} \left| \log \frac{f_{\mu_n}(x)}{f_\mu(x)} \right| < \infty \tag{12}$$

then, $\mu \in \mathcal{H}(\mathbb{X}|\mu_n)$ for all $n \geq 1$, and

$$\lim_{n\to\infty} D(\mu||\mu_n) = 0.$$

Furthermore, $\lim_{n\to\infty} H(\mu_n) = H(\mu)$, if and only if,

$$\lim_{n\to\infty} \sum_{x \in A_{\mu_n} \setminus A_\mu} f_{\mu_n}(x) \log \frac{1}{f_{\mu_n}(x)} = 0. \tag{13}$$

This result shows the non-sufficiency of the convergence in direct I-divergence to achieve entropy convergence in the regime when $\mu \ll \mu_n$. In fact, Lemma 3 may be interpreted as an extension of Lemma 1 when the finite support assumption over μ is relaxed.

3. Shannon Entropy Estimation

Let μ be a probability in $\mathcal{H}(\mathbb{X})$, and let denote by $X_1, X_2, X_3, \ldots$ the empirical process induced from i.i.d. realizations of a random variable driven by μ, i.e., $X_i \sim \mu$, for all $i \geq 0$. Let $\mathbb{P}_\mu$ denote the distribution of the empirical process in $(\mathbb{X}^\infty, \mathcal{B}(\mathbb{X}^\infty))$ and $\mathbb{P}_\mu^n$ denote the finite block distribution of $X^n \equiv (X_1, \ldots, X_n)$ in the product space $(\mathbb{X}^n, \mathcal{B}(\mathbb{X}^n))$. Given a realization of $X_1, X_2, X_3, \ldots, X_n$, we can construct an histogram-based estimator like classical empirical probability given by:

$$\hat{\mu}_n(A) \equiv \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_A(X_k), \ \forall A \subset \mathbb{X}, \tag{14}$$

with pmf given by $f_{\hat{\mu}_n}(x) = \hat{\mu}_n(\{x\})$ for all $x \in \mathbb{X}$. A natural estimator of the entropy is the plug-in estimate of $\hat{\mu}_n$ given by

$$H(\hat{\mu}_n) = -\sum_{x\in\mathbb{X}} f_{\hat{\mu}_n}(x) \log f_{\hat{\mu}_n}(x), \tag{15}$$

which is a measurable function of $X_1, \ldots, X_n$ (this dependency on the data will be implicit for the rest of the paper).

For the rest of this Section and Sections 4 and 5, the convergence results in Section 2.1 are used to derive strong consistency results for plug-in histogram-based estimators, like $H(\hat{\mu}_n)$ in (15), as well as finite length concentration inequalities to obtain almost-sure rates of convergence for the overall estimation error $|H(\hat{\mu}_n) - H(\mu)|$.

3.1. Revisiting the Classical Plug-In Estimator for Finite and Unknown Supported Distributions

We start by analyzing the case when μ has a finite but unknown support. A consequence of the strong law of large numbers [42,43] is that $\forall x \in \mathbb{X}$, $\lim_{n\to\infty} \hat{\mu}_n(\{x\}) = \mu(\{x\})$, $\mathbb{P}_\mu$-almost surely (a.s.), hence $\lim_{n\to\infty} V(\hat{\mu}_n, \mu) = 0$, $\mathbb{P}_\mu$-a.s. On the other hand, it is clear that $A_{\hat{\mu}_n} \subset A_\mu$ holds with probability one. Then Proposition 1 implies that

$$\lim_{n\to\infty} D(\hat{\mu}_n||\mu) = 0 \text{ and } \lim_{n\to\infty} H(\hat{\mu}_n) = H(\mu), \ \mathbb{P}_\mu\text{-a.s.}, \tag{16}$$

i.e., $\hat{\mu}_n$ is a strongly consistent estimator of μ in reverse I-divergence and $H(\hat{\mu}_n)$ is a strongly consistent estimate of $H(\mu)$ distribution-free in $\mathcal{F}(\mathbb{X})$. Furthermore, the following can be stated:

Theorem 1. *Let $\mu \in \mathcal{F}(\mathbb{X})$ and let us consider $\hat{\mu}_n$ in (14). Then $\hat{\mu}_n \in \mathcal{H}(\mathbb{X}) \cap \mathcal{H}(\mathbb{X}|\mu)$, $\mathbb{P}_\mu$-a.s and $\forall n \geq 1$, $\forall \epsilon > 0$,*

$$\mathbb{P}_\mu^n \left(D(\hat{\mu}_n||\mu) > \epsilon\right) \leq 2^{|A_\mu|+1} \cdot e^{-\frac{2\mathbf{m}_\mu^2 \cdot n\epsilon^2}{\log e^2}}, \tag{17}$$

$$\mathbb{P}_\mu^n \left(|H(\hat{\mu}_n) - H(\mu)| > \epsilon\right) \leq 2^{|A_\mu|+1} \cdot e^{-\frac{2n\epsilon^2}{(\mathbf{M}_\mu + \frac{\log e}{\mathbf{m}_\mu})^2}}. \tag{18}$$

Moreover, $D(\mu||\hat{\mu}_n)$ is eventually finite with probability one and $\forall \epsilon > 0$, and for any $n \geq 1$,

$$\mathbb{P}_\mu^n \left(D(\mu||\hat{\mu}_n) > \epsilon\right) \leq 2^{|A_\mu|+1} \cdot \left[e^{-\frac{2n\epsilon^2}{\log e^2 \cdot (1/\mathbf{m}_\mu + 1)^2}} + e^{-n\mathbf{m}_\mu^2}\right]. \tag{19}$$

This result implies that for any $\tau \in (0, 1/2)$ and $\mu \in \mathcal{F}(\mathbb{X})$, $|H(\hat{\mu}_n) - H(\mu)|$, $D(\hat{\mu}_n||\mu)$, and $D(\mu||\hat{\mu}_n)$ goes to zero as $o(n^{-\tau})$ $\mathbb{P}_\mu$-a.s. Furthermore, $\mathbb{E}_{\mathbb{P}_\mu^n}(|H(\hat{\mu}_n) - H(\mu)|)$ and $\mathbb{E}_{\mathbb{P}_\mu^n}(D(\hat{\mu}_n||\mu))$ behave like $O(1/\sqrt{n})$ for all $\mu \in \mathcal{F}(\mathbb{X})$ from (30) in Section 7, which is the optimal rate of convergence of the finite alphabet scenario. As a corollary of (18), it is possible to derive intervals of confidence for the estimation error $|H((\hat{\mu}_n) - H(\mu_n)|$: for all $\delta > 0$ and $n \geq 1$,

$$\mathbb{P}_\mu \left(|H((\hat{\mu}_n) - H(\mu_n)| \leq \left(\mathbf{M}_\mu + \log e/\mathbf{m}_\mu\right) \sqrt{\frac{1}{2n} \ln \frac{2^{|A_\mu|+1}}{\delta}} \right) \geq 1 - \delta. \tag{20}$$

This confidence interval behaves like $O(1/\sqrt{n})$ as a function of n, and like $O(\sqrt{\ln 1/\delta})$ as a function of δ, which are the same optimal asymptotic trends that can be obtained for $V(\mu, \hat{\mu}_n)$ in (30).

Finally, we observe that $A_{\hat{\mu}_n} \subset A_\mu$ $\mathbb{P}_\mu^n$-a.s. where for any $n \geq 1$, $\mathbb{P}_\mu^n(A_{\hat{\mu}_n} \neq A_\mu) > 0$ implying that $\mathbb{E}_{\mathbb{P}_\mu^n}(D(\mu||\hat{\mu}_n)) = \infty$ for all finite n. Then even in the finite and unknown supported scenario, $\hat{\mu}_n$ is not consistent in expected direct I-divergence, which is congruent with the result in [29,44]. Besides this

negative result, strong consistency in direct I-divergence can be obtained from (19), in the sense that $\lim_{n\to\infty} D(\mu||\hat{\mu}_n) = 0$, $\mathbb{P}_\mu$-a.s.

3.2. A Simplified Version of the Barron Estimator for Finite Supported Probabilities

It is well-understood that consistency in expected direct I-divergence is of critical importance for the construction of a lossless universal source coding scheme [2,23,29,44–48]. Here, we explore an estimator that achieves this learning objective, in addition to entropy estimation. For that, let $\mu \in \mathcal{F}(\mathbb{X})$ and let assume $v \in \mathcal{F}(\mathbb{X})$ such that $\mu \ll v$. Barron et al. [29] proposed a modified version of the empirical measure in (14) to estimate μ from i.i.d. realizations, adopting a mixture estimate of the form

$$\tilde{\mu}_n(B) = (1 - a_n) \cdot \hat{\mu}_n(B) + a_n \cdot v(B), \tag{21}$$

for all $B \subset \mathbb{X}$, and with $(a_n)_{n\in\mathbb{N}}$ a sequence of real numbers in $(0,1)$. Note that $A_{\tilde{\mu}_n} = A_v$ then $\mu \ll \tilde{\mu}_n$ for all n and from the finite support assumption $H(\tilde{\mu}_n) < \infty$ and $D(\mu||\tilde{\mu}_n) < \infty$, $\mathbb{P}_\mu$-a.s.. The following result derives from the convergence result in Lemma 1.

Theorem 2. *Let $v \in \mathcal{F}(\mathbb{X})$, $\mu \ll v$ and let us consider $\tilde{\mu}_n$ in (21) induced from i.i.d. realizations of μ.*

(i) If (a_n) is $o(1)$, then $\lim_{n\to\infty} H(\tilde{\mu}_n) = H(\mu)$, $\lim_{n\to\infty} D(\mu||\tilde{\mu}_n) = 0$, $\mathbb{P}_\mu$-a.s., and $\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_\mu}(D(\mu||\tilde{\mu}_n)) = 0$.

(ii) Furthermore, if (a_n) is $O(n^{-p})$ with $p > 2$, then for all $\tau \in (0, 1/2)$, $|H(\tilde{\mu}_n) - H(\mu)|$ and $D(\mu||\tilde{\mu}_n)$ are $o(n^{-\tau})$ $\mathbb{P}_\mu$-a.s, and $\mathbb{E}_{\mathbb{P}_\mu}(|H(\tilde{\mu}_n) - H(\mu)|)$ and $\mathbb{E}_{\mathbb{P}_\mu}(D(\mu||\tilde{\mu}_n))$ are $O(1/\sqrt{n})$.

Using this approach, we achieve estimation of the true distribution in expected information divergence as well as strong consistency for entropy estimation as intended. In addition, optimal rates of convergence are obtained under the finite support assumption on μ.

4. The Barron-Györfi-van der Meulen Estimator

The celebrated Barron estimator was proposed by Barron, Györfi and van der Meulen [29] in the context of an abstract and continuous measurable space. It is designed as a variation of the classical histogram-based scheme to achieve a consistent estimation of the distribution in direct I-divergence [29] (Theorem 2). Here, the Barron estimator is revisited in the countable alphabet scenario, with the objective of estimating the Shannon entropy consistently, which, to the best of our knowledge, has not been previously addressed in literature. For that purpose, the convergence result in Lemma 3 will be used as a key result.

Let $v \in \mathcal{P}(\mathbb{X})$ be of infinite support (i.e., $\mathbf{m}_v = \inf_{x\in A_v} v(\{x\}) = 0$). We want to construct a strongly consistent estimate of the entropy restricted to the collection of probabilities in $\mathcal{H}(\mathbb{X}|v)$. For that, let us consider a sequence $(h_n)_{n\geq 0}$ with values in $(0,1)$ and let us denote by $\pi_n = \{A_{n,1}, A_{n,2}, \ldots, A_{n,m_n}\}$ the finite partition of $\mathbb{X}$ with maximal cardinality satisfying that

$$v(A_{n,i}) \geq h_n, \ \forall i \in \{1, \ldots, m_n\}. \tag{22}$$

Note that $m_n = |\pi_n| \leq 1/h_n$ for all $n \geq 1$, and because of the fact that $\inf_{x\in A_v} v(\{x\}) = 0$ it is simple to verify that if (h_n) is $o(1)$ then $\lim_{n\to\infty} m_n = \infty$. π_n offers an approximated statistically equivalent partition of $\mathbb{X}$ with respect to the reference measure v. In this context, given $X_1, \ldots, X_n$, i.i.d. realizations of $\mu \in \mathcal{H}(\mathbb{X}|v)$, the idea proposed by Barron et al. [29] is to estimate the RN derivative $\frac{d\mu}{dv}(x)$ by the following histogram-based construction:

$$\frac{d\mu_n^*}{dv}(x) = (1 - a_n) \cdot \frac{\hat{\mu}_n(A_n(x))}{v(A_n(x))} + a_n, \ \forall x \in A_v, \tag{23}$$

where a_n is a real number in $(0,1)$, $A_n(x)$ denotes the cell in π_n that contains the point x, and $\hat{\mu}_n$ is the empirical measure in (14). Note that

$$f_{\mu_n^*}(x) = \frac{d\mu_n^*}{d\lambda}(x) = f_v(x) \cdot \left[(1-a_n) \cdot \frac{\hat{\mu}_n(A_n(x))}{v(A_n(x))} + a_n \right],$$

$\forall x \in \mathbb{X}$, and, consequently, $\forall B \subset \mathbb{X}$

$$\mu_n^*(B) = (1-a_n) \sum_{i=1}^{m_n} \hat{\mu}_n(A_{n,i}) \cdot \frac{v(B \cap A_{n,i})}{v(A_{n,i})} + a_n v(B). \tag{24}$$

By construction $A_\mu \subset A_v \subset A_{\mu_n^*}$ and, consequently, $\mu \ll \mu_n^*$ for all $n \geq 1$. The next result shows sufficient conditions on the sequences (a_n) and (h_n) to guarantee a strongly consistent estimation of the entropy $H(\mu)$ and of μ in direct I-divergence, distribution free in $\mathcal{H}(\mathbb{X}|v)$. The proof is based on verifying that the sufficient conditions of Lemma 3 are satisfied $\mathbb{P}_\mu$-a.s.

Theorem 3. *Let v be in $\mathcal{P}(\mathbb{X}) \cap \mathcal{H}(\mathbb{X})$ with infinite support, and let us consider μ in $\mathcal{H}(\mathbb{X}|v)$. If we have that:*

(i) *(a_n) is $o(1)$ and (h_n) is $o(1)$,*
(ii) *$\exists \tau \in (0, 1/2)$, such that the sequence $\left(\frac{1}{a_n \cdot h_n}\right)$ is $o(n^\tau)$,*

then $\mu \in \mathcal{H}(\mathbb{X}) \cap \mathcal{H}(\mathbb{X}|\mu_n^*)$ for all $n \geq 1$ and

$$\lim_{n\to\infty} H(\mu_n^*) = H(\mu) \text{ and } \lim_{n\to\infty} D(\mu||\mu_n^*) = 0, \ \mathbb{P}_\mu\text{-a.s..} \tag{25}$$

This result shows an admisible regime of design parameters and its scaling with the number of samples that guarantees that the Barron plug-in entropy estimator is strongly consistent in $\mathcal{H}(\mathbb{X}|v)$. As a byproduct, we obtain that the distribution μ is estimated consistently in direct information divergence.

The Barron estimator [29] was originally proposed in the context of distributions defined in an abstract measurable space. Then if we restrict [29] (Theorem 2) to the countable alphabet case, the following result is obtained:

Corollary 1. *Ref. [29] (Theorem 2) Let us consider $v \in \mathcal{P}(\mathbb{X})$ and $\mu \in \mathcal{H}(\mathbb{X}|v)$. If (a_n) is $o(1)$, (h_n) is $o(1)$ and $\limsup_{n\to\infty} \frac{1}{n a_n h_n} \leq 1$ then*

$$\lim_{n\to\infty} D(\mu||\mu_n^*) = 0, \ \mathbb{P}_\mu\text{-a.s.}$$

When the only objective is the estimation of distributions consistently in direct I-divergence, Corollary 1 should be considered to be a better result than Theorem 3 (Corollary 1 offers weaker conditions than Theorem 3 in particular condition (ii)). The proof of Theorem 3 is based on verifying the sufficient conditions of Lemma 3, where the objective is to achieve the convergence of the entropy, and as a consequence, the convergence in direct I-divergence. Therefore, we can say that the stronger conditions of Theorem 3 are needed when the objective is entropy estimation. This is justified from the observation that convergence in direct I-divergence does not imply entropy convergence in ∞-alphabets, as is discussed in Section 2.1 (see, Lemmas 1 and 3).

5. A Data-Driven Histogram-Based Estimator

Data-driven partitions offer a better approximation to the data distribution in the sample space than conventional non-adaptive histogram-based approaches [34,49]. They have the capacity to improve the approximation quality of histogram-based learning schemes, which translates in better performance in different non-parametric learning settings [34–36,50,51]. One of the basic design principles of this approach is to partition or select a sub-set of elements of $\mathbb{X}$ in a data-dependent

way to preserve a critical number of samples per cell. In our problem, this last condition proves to be crucial to derive bounds for the estimation and approximation errors. Finally, these expressions will be used to propose design solutions that offer an optimal balance between estimation and approximation errors (Theorems 5 and 6).

Given $X_1, \ldots, X_n$ i.i.d. realizations driven by $\mu \in \mathcal{H}(\mathbb{X})$ and $\epsilon > 0$, let us define the data-driven set

$$\Gamma_\epsilon \equiv \{x \in \mathbb{X} : \hat{\mu}_n(\{x\}) \geq \epsilon\}, \tag{26}$$

and $\phi_\epsilon \equiv \Gamma_\epsilon^c$. Let $\Pi_\epsilon \equiv \{\{x\} : x \in \Gamma_\epsilon\} \cup \{\phi_\epsilon\} \subset 2^{\mathbb{X}}$ be a data-driven partition with maximal resolution in Γ_ϵ, and $\sigma_\epsilon \equiv \sigma(\Pi_\epsilon)$ be the smallest sigma field that contains Π_ϵ (Π_ϵ is a finite partition, σ_ϵ is the collection of sets that are union of elements of Π_ϵ). We propose the conditional empirical probability restricted to Γ_ϵ by:

$$\hat{\mu}_{n,\epsilon} \equiv \hat{\mu}_n(\cdot|\Gamma_\epsilon). \tag{27}$$

By construction, it follows that $A_{\hat{\mu}_{n,\epsilon}} = \Gamma_\epsilon \subset A_\mu$, $\mathbb{P}_\mu$-a.s. and this implies that $\hat{\mu}_{n,\epsilon} \ll \mu$ for all $n \geq 1$. Furthermore, $|\Gamma_\epsilon| \leq \frac{1}{\epsilon}$ and, importantly in the context of the entropy functional, it follows that

$$\mathbf{m}_{\hat{\mu}_n}^{\epsilon} \equiv \inf_{x \in \Gamma_\epsilon} \hat{\mu}_n(\{x\}) \geq \epsilon. \tag{28}$$

The next result establishes a mild sufficient condition on (ϵ_n) for which $H(\hat{\mu}_{n,\epsilon_n})$ is strongly consistent distribution-free in $\mathcal{H}(\mathbb{X})$. Considering that we are in the regime where $\hat{\mu}_{n,\epsilon_n} \ll \mu$, $\mathbb{P}_\mu$-a.s., the proof of this result uses the convergence result in Lemma 2 as a central result.

Theorem 4. *If (ϵ_n) is $O(n^{-\tau})$ with $\tau \in (0,1)$, then for all $\mu \in \mathcal{H}(\mathbb{X})$*

$$\lim_{n\to\infty} H(\hat{\mu}_{n,\epsilon_n}) = H(\mu), \ \mathbb{P}_\mu\text{-a.s.}$$

Complementing Theorem 4, the next result offers almost-sure rates of converge for a family of distributions with a power tail bounded condition (TBC). In particular, the family of distributions studied in [30] (Theorem 7) are considered.

Theorem 5. *Let us assume that for some $p > 1$ there are two constants $0 < k_0 \leq k_1$ and $N > 0$ such that $k_0 \cdot x^{-p} \leq \mu(\{x\}) \leq k_1 x^{-p}$ for all $x \geq N$. If we consider that $(\epsilon_n) \approx (n^{-\tau^*})$ for $\tau^* = \frac{1}{2+1/p}$, then*

$$|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})| \text{ is } O(n^{-\frac{1-1/p}{2+1/p}}), \ \mathbb{P}_\mu\text{-a.s.}$$

This result shows that under the mentioned p-power TBC on $f_\mu(\cdot)$, the plug-in estimator $H(\hat{\mu}_{n,\epsilon_n})$ can achieve a rate of convergence to the true limit that is $O(n^{-\frac{1-1/p}{2+1/p}})$ with probability one. For the derivation of this result, the approximation sequence (ϵ_n) is defined as a function of p (adapted to the problem) by finding an optimal tradeoff between estimation and approximation errors while performing a finite length (non-asymptotic) analysis of the expression $|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})|$ (the details of this analysis are presented in Section 7).

It is insightful to look at two extreme regimes of this result: p approaching 1, in which the rate is arbitrarily slow (approaching a non-decaying behavior); and $p \to \infty$, where $|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})|$ is $O(n^{-q})$ for all $q \in (0, 1/2)$ $\mathbb{P}_\mu$-a.s.. This last power decaying range $q \in (0, 1/2)$ matches what is achieved for the finite alphabet scenario (for instance in Theorem 1, Equation (18)), which is known to be the optimal rate for finite alphabets.

Extending Theorem 5, the following result addresses the more constrained case of distributions with an exponential TBC.

Theorem 6. *Let us consider $\alpha > 0$ and let us assume that there are k_0, k_1 with $0 < k_0 \leq k_1$ and $N > 0$ such that $k_0 \cdot e^{-\alpha x} \leq \mu(\{x\}) \leq k_1 \cdot e^{-\alpha x}$ for all $x \geq N$. If we consider $(\epsilon_n) \approx (n^{-\tau})$ with $\tau \in (0, 1/2)$, then*

$$|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})| \text{ is } O(n^{-\tau} \log n), \ \mathbb{P}_\mu\text{-a.s.}$$

Under this stringent TBC on $f_\mu(\cdot)$, it is observed that $|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})|$ is $o(n^{-q})$ $\mathbb{P}_\mu$-a.s., for any arbitrary $q \in (0, 1/2)$, by selecting $(\epsilon_n) \approx (n^{-\tau})$ with $q < \tau < 1/2$. This last condition on τ is universal over $\alpha > 0$. Remarkably, for any distribution with this exponential TBC, we can approximate (arbitrarily closely) the optimal almost-sure rate of convergence achieved for the finite alphabet problem.

Finally, the finite and unknown supported scenario is revisited, where it is shown that the data-driven estimator exhibits the optimal almost sure convergence rate of the classical plug-in entropy estimator presented in Section 3.1.

Theorem 7. *Let us assume that $\mu \in \mathcal{F}(\mathbb{X})$ and (ϵ_n) being $o(1)$. Then for all $\epsilon > 0$ there is $N > 0$ such that $\forall n \geq N$*

$$\mathbb{P}_\mu^n \left(|H(\hat{\mu}_{n,\epsilon_n}) - H(\mu)| > \epsilon \right) \leq 2^{|A_\mu|+1} \cdot \left[e^{-\frac{2n\epsilon^2}{(\mathbf{M}_\mu + \frac{\log e}{\mathbf{m}_\mu})^2}} + e^{-\frac{n\mathbf{m}_\mu^2}{4}} \right]. \tag{29}$$

The proof of this result reduces to verify that $\hat{\mu}_{n,\epsilon_n}$ detects A_μ almost-surely when n goes to infinity and from this, it follows that $H(\hat{\mu}_{n,\epsilon_n})$ eventually matches the optimal almost sure performance of $H(\hat{\mu}_n)$ under the key assumption that $\mu \in \mathcal{F}(\mathbb{X})$. Finally, the concentration bound in (29) implies that $|H(\hat{\mu}_{n,\epsilon_n}) - H(\mu)|$ is $o(n^{-q})$ almost surely for all $q \in (0, 1/2)$ as long as $\epsilon_n \to 0$ with n.

6. Discussion of the Results and Final Remarks

This work shows that entropy convergence results are instrumental to derive new (strongly consistent) estimation results for the Shannon entropy in ∞-alphabets and, as a byproduct, distribution estimators that are strongly consistent in direct and reverse I-divergence. Adopting a set of sufficient conditions for entropy convergence in the context of four plug-in histogram-based schemes, this work shows concrete design conditions where strong consistency for entropy estimation in ∞-alphabets can be obtained (Theorems 2–4). In addition, the relevant case where the target distribution has a finite but unknown support is explored, deriving almost sure rates of convergence results for the overall estimation error (Theorems 1 and 7) that match the optimal asymptotic rate that can be obtained in the finite alphabet version of the problem (i.e., the finite and known supported case).

As the main context of application, this work focuses on the case of a data-driven plug-in estimator that restricts the support where the distribution is estimated. The idea is to have design parameters that control estimation-error effects and to find an adequate balance between these two learning errors. Adopting the entropy convergence result in Lemma 2, it is shown that this data-driven scheme offers the same universal estimation attributes than the classical plug-in estimate under some mild conditions on its threshold design parameter (Theorem 4). In addition, by addressing the technical task of deriving concrete closed-form expressions for the estimation and approximation errors in this learning context a solution is presented where almost-sure rates of convergence of the overall estimation error are obtained over a family of distributions with some concrete tail bounded conditions (Theorems 5 and 6). These results show the capacity that data-driven frameworks offer for adapting aspects of their learning scheme to the complexity of the entropy estimation task in ∞-alphabets.

Concerning the classical plug-in estimator presented in Section 3.1, it is important to mention that the work of Antos et al. [30] shows that $\lim_{n\to\infty} H(\hat{\mu}_n) = H(\mu)$ happens almost surely and distribution-free and, furthermore, it provides rates of convergence for families with specific tail-bounded conditions [30] (Theorem 7). Theorem 1 focuses on the case when $\mu \in \mathcal{F}(\mathbb{X})$, where new finite-length deviation inequalities and confidence intervals are derived. From that perspective, Theorem 1 complements the

results presented in [30] in the non-explored scenario when $\mu \in \mathcal{F}(\mathbb{X})$. It is also important to mention two results by Ho and Yeung [26] (Theorems 11 and 12) for the plug-in estimator in (15). They derived bounds for $\mathbb{P}_\mu^n(|H(\hat{\mu}_n) - H(\mu)| \geq \epsilon)$ and determined confidence intervals under a finite and known support restriction on μ. In contrast, Theorem 1 resolves the case for a finite and unknown supported distribution, which is declared to be a challenging problem from the arguments presented in [26] (Theorem 13) concerning the discontinuity of the entropy.

7. Proof of the Main Results

Proof of Theorem 1. Let μ be in $\mathcal{F}(\mathbb{X})$, then $\left|A_\mu\right| \leq k$ for some $k > 1$. From Hoeffding's inequality [28], $\forall n \geq 1$, and for any $\epsilon > 0$

$$\mathbb{P}_\mu^n\left(V(\hat{\mu}_n, \mu) > \epsilon\right) \leq 2^{k+1} \cdot e^{-2n\epsilon^2} \text{ and } \mathbb{E}_{\mathbb{P}_\mu^n}(V(\hat{\mu}_n, \mu)) \leq 2\sqrt{\frac{(k+1)\log 2}{n}}. \tag{30}$$

Considering that $\hat{\mu}_n \ll \mu$ $\mathbb{P}_\mu$-a.s, we can use Proposition 1 to obtain that

$$D(\hat{\mu}_n||\mu) \leq \frac{\log e}{\mathbf{m}_\mu} \cdot V(\hat{\mu}_n, \mu), \text{ and } |H((\hat{\mu}_n) - H(\mu_n)| \leq \left[\mathbf{M}_\mu + \frac{\log e}{\mathbf{m}_\mu}\right] \cdot V(\hat{\mu}_n, \mu). \tag{31}$$

Hence, (17) and (18) derive from (30).

For the direct I-divergence, let us consider a sequence $(x_i)_{i\geq 1}$ and the following function (a stopping time):

$$T_o(x_1, x_2, \ldots) \equiv \inf\left\{n \geq 1 : A_{\hat{\mu}_n(x^n)} = A_\mu\right\}. \tag{32}$$

$T_o(x_1, x_2, \ldots)$ is the point where the support of $\hat{\mu}_n(x^n)$ is equal to A_μ and, consequently, the direct I-divergence is finite (since $\mu \in \mathcal{F}(\mathbb{X})$). In fact, by the uniform convergence of $\hat{\mu}_n$ to μ_n ($\mathbb{P}_\mu$-a.s.) and the finite support assumption of μ, it is simple to verify that $\mathbb{P}_\mu(T_o(X_1, X_2, \ldots) < \infty) = 1$. Let us define the event:

$$\mathcal{B}_n \equiv \{x_1, x_2, \cdots : T_o(x_1, x_2, \ldots) \leq n\} \subset \mathbb{X}^{\mathbb{N}}, \tag{33}$$

i.e., the collection of sequences in $\mathbb{X}^{\mathbb{N}}$ where at time n, $A_{\hat{\mu}_n} = A_\mu$ and, consequently, $D(\mu||\hat{\mu}_n) < \infty$. Restricted to this set

$$\begin{aligned} D(\mu||\hat{\mu}_n) &\leq \sum_{x \in A_{\hat{\mu}_n||\mu}} f_{\hat{\mu}_n}(x) \log \frac{f_{\hat{\mu}_n}(x)}{f_\mu(x)} + \sum_{x \in A_\mu \setminus A_{\hat{\mu}_n||\mu}} f_{\hat{\mu}_n}(x) \log \frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)} && (34) \\ &\leq \log e \cdot \sum_{x \in A_{\hat{\mu}_n||\mu}} f_{\hat{\mu}_n}(x) \cdot \left(\frac{f_{\hat{\mu}_n}(x)}{f_\mu(x)} - 1\right) \\ &+ \log e \cdot \left[\mu(A_\mu \setminus A_{\hat{\mu}_n||\mu}) - \hat{\mu}_n((A_\mu \setminus A_{\hat{\mu}_n||\mu}))\right] && (35) \\ &\leq \log e \cdot (1/\mathbf{m}_u + 1)\, V(\mu, \hat{\mu}_n), && (36) \end{aligned}$$

where in the first inequality $A_{\hat{\mu}_n||\mu} \equiv \left\{x \in A_{\hat{\mu}_n} : f_{\hat{\mu}_n}(x) > f_\mu(x)\right\}$, and the last is obtained by the definition of the total variational distance. In addition, let us define the ϵ-deviation set $\mathcal{A}_n^\epsilon \equiv \{x_1, x_2, \cdots : D(\mu||\hat{\mu}_n(x^n)) > \epsilon\} \subset \mathbb{X}^{\mathbb{N}}$. Then by additivity and monotonicity of $\mathbb{P}_\mu$, we have that

$$\mathbb{P}_\mu(\mathcal{A}_n^\epsilon) \leq \mathbb{P}_\mu(\mathcal{A}_n^\epsilon \cap \mathcal{B}_n) + \mathbb{P}_\mu(\mathcal{B}_n^c). \tag{37}$$

By definition of $\mathcal{B}_n$, (36) and (30) it follows that

$$\begin{aligned}\mathbb{P}_\mu(\mathcal{A}_n^\epsilon \cap \mathcal{B}_n) &\leq \mathbb{P}_\mu(V(\mu||\hat{\mu}_n)\log e \cdot (1/\mathbf{m}_u + 1) > \epsilon) \\ &\leq 2^{|A_\mu|+1} \cdot e^{-\frac{2n\epsilon^2}{\log e^2 \cdot (1/\mathbf{m}_u+1)^2}}. \end{aligned} \tag{38}$$

On the other hand, $\forall \epsilon_o \in (0, \mathbf{m}_\mu)$ if $V(\mu, \hat{\mu}_n) \leq \epsilon_o$ then $T_o \leq n$. Consequently $\mathcal{B}_n^c \subset \{x_1, x_2, \cdots : V(\mu, \hat{\mu}_n(x^n)) > \epsilon_o\}$, and again from (30)

$$\mathbb{P}_\mu(\mathcal{B}_n^c) \leq 2^{|A_\mu|+1} \cdot e^{-2n\epsilon_o^2}, \tag{39}$$

for all $n \geq 1$ and $\forall \epsilon_o \in (0, \mathbf{m}_\mu)$. Integrating the results in (38) and (39) and considering $\epsilon_0 = \mathbf{m}_\mu/\sqrt{2}$ suffice to show the bound in (19). □

Proof of Theorem 2. As (a_n) is $o(1)$, it is simple to verify that $\lim_{n\to\infty} V(\tilde{\mu}_n, \mu) = 0$, $\mathbb{P}_\mu$-a.s. Also note that the support disagreement between $\tilde{\mu}_n$ and μ is bounded by the hypothesis, then

$$\lim_{n\to\infty} \tilde{\mu}_n\left(A_{\mu_n} \setminus A_\mu\right) \cdot \log\left|A_{\tilde{\mu}_n} \setminus A_\mu\right| \leq \lim_{n\to\infty} \tilde{\mu}_n\left(A_{\mu_n} \setminus A_\mu\right) \cdot \log|A_v| = 0, \ \mathbb{P}_\mu\text{-a.s.} \tag{40}$$

Therefore from Lemma 1, we have the strong consistency of $H(\tilde{\mu}_n)$ and the almost sure convergence of $D(\mu||\tilde{\mu}_n)$ to zero. Note that $D(\mu||\tilde{\mu}_n)$ is uniformly upper bounded by $\log e \cdot (1/\mathbf{m}_\mu + 1)V(\mu, \tilde{\mu}_n)$ (see (36) in the proof of Theorem 1). Then the convergence in probability of $D(\mu||\tilde{\mu}_n)$ implies the convergence of its mean [42], which concludes the proof of the first part.

Concerning rates of convergence, we use the following:

$$\begin{aligned}H(\mu) - H(\tilde{\mu}_n) = &\sum_{x\in A_\mu \cap A_{\tilde{\mu}_n}} \left[f_{\tilde{\mu}_n}(x) - f_\mu(x)\right]\log f_\mu(x) + \sum_{x\in A_\mu\cap A_{\tilde{\mu}_n}} f_{\tilde{\mu}_n}(x)\log\frac{f_{\tilde{\mu}_n}(x)}{f_\mu(x)} \\ &- \sum_{x\in A_{\tilde{\mu}_n}\setminus A_\mu} f_{\tilde{\mu}_n}(x)\log\frac{1}{f_{\tilde{\mu}_n}(x)}. \end{aligned} \tag{41}$$

The absolute value of the first term in the right hand side (RHS) of (41) is bounded by $\mathbf{M}_\mu \cdot V(\tilde{\mu}_n, \mu)$ and the second term is bounded by $\log e/\mathbf{m}_\mu \cdot V(\tilde{\mu}_n, \mu)$, from the assumption that $\mu \in \mathcal{F}(\mathbb{X})$. For the last term, note that $f_{\tilde{\mu}_n}(x) = a_n \cdot v(\{x\})$ for all $x \in A_{\tilde{\mu}_n} \setminus A_\mu$ and that $A_{\tilde{\mu}_n} = A_v$, then

$$0 \leq \sum_{x\in A_{\tilde{\mu}_n}\setminus A_\mu} f_{\tilde{\mu}_n}(x)\log\frac{1}{f_{\tilde{\mu}_n}(x)} \leq a_n \cdot (H(v) + \log\frac{1}{a_n} \cdot v(A_v \setminus A_\mu)).$$

On the other hand,

$$\begin{aligned}V(\tilde{\mu}_n, \mu) &= \frac{1}{2}\sum_{x\in A_\mu} |(1-a_n)\hat{\mu}_n(\{x\}) + a_n v(\{x\}) - \mu(\{x\})| + \sum_{x\in A_v\setminus A_\mu} a_n v(\{x\}). \\ &\leq (1-a_n)\cdot V(\hat{\mu}_n, \mu) + a_n.\end{aligned}$$

Integrating these bounds in (41),

$$\begin{aligned}|H(\mu) - H(\tilde{\mu}_n)| &\leq (\mathbf{M}_\mu + \log e/\mathbf{m}_\mu)\cdot((1-a_n)\cdot V(\hat{\mu}_n, \mu) + a_n) + a_n \cdot H(v) + a_n \cdot \log\frac{1}{a_n} \\ &= K_1 \cdot V(\hat{\mu}_n, \mu) + K_2 \cdot a_n + a_n \cdot \log\frac{1}{a_n}, \end{aligned} \tag{42}$$

for constants $K_1 > 0$ and $K_2 > 0$ function of μ and v.

Under the assumption that $\mu \in \mathcal{F}(\mathbb{X})$, the Hoeffding's inequality [28,52] tells us that $\mathbb{P}_\mu(V(\hat{\mu}_n,\mu) > \epsilon) \leq C_1 \cdot e^{-C_2 n\epsilon^2}$ (for some distribution free constants $C_1 > 0$ and $C_2 > 0$). From this inequality, $V(\hat{\mu}_n,\mu)$ goes to zero as $o(n^{-\tau})$ $\mathbb{P}_\mu$-a.s. $\forall \tau \in (0,1/2)$ and $\mathbb{E}_{\mathbb{P}_\mu}(V(\hat{\mu}_n,\mu))$ is $O(1/\sqrt{n})$. On the other hand, under the assumption in ii) $(K_2 \cdot a_n + a_n \cdot \log \frac{1}{a_n})$ is $O(1/\sqrt{n})$, which from (42) proves the rate of convergence results for $|H(\mu) - H(\tilde{\mu}_n)|$.

Considering the direct I-divergence, $D(\mu||\tilde{\mu}_n) \leq \log e \cdot \sum_{x\in A_\mu} f_\mu(x) \left|\frac{f_\mu(x)}{f_{\tilde{\mu}_n}(x)} - 1\right| \leq \frac{\log e}{\mathbf{m}_{\tilde{\mu}_n}} \cdot V(\tilde{\mu}_n,\mu)$. Then the uniform convergence of $\tilde{\mu}_n(\{x\})$ to $\mu(\{x\})$ $\mathbb{P}_\mu$-a.s. in A_μ and the fact that $\left|A_\mu\right| < \infty$ imply that for an arbitrary small $\epsilon > 0$ (in particular smaller than $\mathbf{m}_\mu$)

$$\lim_{n\longrightarrow\infty} D(\mu||\tilde{\mu}_n) \leq \frac{\log e}{\mathbf{m}_\mu - \epsilon} \cdot \lim_{n\longrightarrow\infty} V(\tilde{\mu}_n,\mu),\ \mathbb{P}_\mu\text{-a.s.}. \tag{43}$$

(43) suffices to obtain the convergence result for the I-divergence. □

Proof of Theorem 3. Let us define the oracle Barron measure $\tilde{\mu}_n$ by:

$$f_{\tilde{\mu}_n}(x) = \frac{d\tilde{\mu}_n}{d\lambda}(x) = f_v(x)\left[(1-a_n)\cdot\frac{\mu(A_n(x))}{v(A_n(x))} + a_n\right], \tag{44}$$

where we consider the true probability instead of its empirical version in (23). Then, the following convergence result can be obtained (see Proposition A2 in Appendix B),

$$\lim_{n\to\infty}\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{d\tilde{\mu}_n}{d\mu_n^*}(x) - 1\right| = 0,\ \mathbb{P}_\mu\text{-a.s.}. \tag{45}$$

Let $\mathcal{A}$ denote the collection of sequences $x_1, x_2, \ldots$ where the convergence in (45) is holding (this set is typical meaning that $\mathbb{P}_\mu(\mathcal{A}) = 1$). The rest of the proof reduces to show that for any arbitrary $(x_n)_{n\geq 1} \in \mathcal{A}$, its respective sequence of induced measures $\{\mu_n^* : n \geq 1\}$ (the dependency of μ_n^* on the sequence $(x_n)_{n\geq 1}$ will be considered implicit for the rest of the proof) satisfies the sufficient conditions of Lemma 3.

Let us fix an arbitrary $(x_n)_{n\geq 1} \in \mathcal{A}$:

Weak convergence $\mu_n^* \Rightarrow \mu$: Without loss of generality we consider that $A_{\tilde{\mu}_n} = A_v$ for all $n \geq 1$. Since $a_n \to 0$ and $h_n \to 0$, $f_{\tilde{\mu}_n}(x) \to \mu(\{x\})$ $\forall x \in A_v$, we got the weak convergence of $\tilde{\mu}_n$ to μ. On the other hand by definition of $\mathcal{A}$, $\lim_{n\to\infty}\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{f_{\tilde{\mu}_n}(x)}{f_{\mu_n^*}(x)} - 1\right| = 0$ that implies that $\lim_{n\to\infty}\left|f_{\mu_n^*}(x) - f_{\tilde{\mu}_n}(x)\right| = 0$ for all $x \in A_v$ and, consequently, $\mu_n^* \Rightarrow \mu$.

The condition in (12): By construction $\mu \ll \mu_n^*$, $\mu \ll \tilde{\mu}_n$ and $\tilde{\mu}_n \approx \mu_n^*$ for all n, then we will use the following equality:

$$\log\frac{d\mu}{d\mu_n^*}(x) = \log\frac{d\mu}{d\tilde{\mu}_n}(x) + \log\frac{d\tilde{\mu}_n}{d\mu_n^*}(x), \tag{46}$$

for all $x \in A_\mu$. Concerning the approximation error term of (46), i.e., $\log\frac{d\mu}{d\tilde{\mu}_n}(x)$, $\forall x \in A_\mu$

$$\frac{d\tilde{\mu}_n}{d\mu}(x) = (1-a_n)\left[\frac{\mu(A_n(x))}{\mu(\{x\})}\frac{v(\{x\})}{v(A_n(x))}\right] + a_n\frac{v(\{x\})}{\mu(\{x\})}. \tag{47}$$

Given that $\mu \in \mathcal{H}(\mathbb{X}|v)$, this is equivalent to state that $\log(\frac{d\mu}{dv}(x))$ is bounded μ-almost everywhere, which is equivalent to say that $\mathbf{m} \equiv \inf_{x\in A_\mu}\frac{d\mu}{dv}(x) > 0$ and $M \equiv \sup_{x\in A_\mu}\frac{d\mu}{dv}(x) < \infty$. From this, $\forall A \subset A_\mu$,

$$\mathbf{m}v(A) \leq \mu(A) \leq Mv(A). \tag{48}$$

Then we have that, $\forall x \in A_\mu$ $\frac{\mathbf{m}}{M} \leq \left[\frac{\mu(A_n(x))}{\mu(\{x\})}\frac{v(\{x\})}{v(A_n(x))}\right] \leq \frac{M}{\mathbf{m}}$. Therefore for n sufficient large, $0 < \frac{1}{2}\frac{\mathbf{m}}{M} \leq \frac{d\tilde{\mu}_n}{d\mu}(x) \leq \frac{M}{\mathbf{m}} + M < \infty$ for all x in A_μ. Hence, there exists $N_o > 0$ such that $\sup_{n\geq N_o}\sup_{x\in A_\mu}\left|\log\frac{d\tilde{\mu}_n}{d\mu}(x)\right| < \infty$.

For the estimation error term of (46), i.e., $\log\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)$, note that from the fact that $(x_n) \in \mathcal{A}$, and the convergence in (45), there exists $N_1 > 0$ such that for all $n \geq N_1$ $\sup_{x\in A_\mu}\left|\log\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)\right| < \infty$, given that $A_\mu \subset A_{\tilde{\mu}_n} = A_v$. Then using (46), for all $n \geq \max\{N_0, N_1\}$ $\sup_{x\in A_\mu}\left|\log\frac{d\mu_n^*}{d\mu}(x)\right| < \infty$, which verifies (12).

The condition in (13): Defining the function $\phi_n^*(x) \equiv 1_{A_v\setminus A_\mu}(x)\cdot f_{\mu_n^*}(x)\log(1/f_{\mu_n^*}(x))$, we want to verify that $\lim_{n\to\infty}\int_{\mathbb{X}}\phi_n^*(x)d\lambda(x) = 0$. Considering that $(x_n) \in \mathcal{A}$ for all $\epsilon > 0$, there exists $N(\epsilon) > 0$ such that $\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{f_{\tilde{\mu}_n}(x)}{f_{\mu_n^*}(x)} - 1\right| < \epsilon$ and then

$$(1-\epsilon)f_{\tilde{\mu}_n}(x) < f_{\mu_n^*}(x) < (1+\epsilon)f_{\tilde{\mu}_n}(x), \text{ for all } x \in A_v. \tag{49}$$

From (49), $0 \leq \phi_n^*(x) \leq (1+\epsilon)f_{\tilde{\mu}_n}(x)\log(1/(1-\epsilon)f_{\tilde{\mu}_n}(x))$ for all $n \geq N(\epsilon)$. Analyzing $f_{\tilde{\mu}_n}(x)$ in (44), there are two scenarios: $A_n(x)\cap A_\mu = \emptyset$ where $f_{\tilde{\mu}_n}(x) = a_n f_v(x)$ and, otherwise, $f_{\tilde{\mu}_n}(x) = f_v(x)(a_n + (1-a_n)\mu(A_n(x)\cap A_\mu)/v(A_n(x)))$. Let us define:

$$\mathcal{B}_n \equiv \left\{x \in A_v \setminus A_\mu : A_n(x)\cap A_\mu = \emptyset\right\} \text{ and } \mathcal{C}_n \equiv \left\{x \in A_v \setminus A_\mu : A_n(x)\cap A_\mu \neq \emptyset\right\}. \tag{50}$$

Then for all $n \geq N(\epsilon)$,

$$\begin{aligned}\sum_{x\in\mathbb{X}}\phi_n^*(x) &\leq \sum_{x\in A_v\setminus A_\mu}(1+\epsilon)f_{\tilde{\mu}_n}(x)\log 1/((1-\epsilon)f_{\tilde{\mu}_n}(x)) \\ &= \sum_{x\in\mathcal{B}_n}(1+\epsilon)a_n f_v(x)\log\frac{1}{(1-\epsilon)a_n f_v(x)} + \sum_{x\in\mathbb{X}}\tilde{\phi}_n(x),\end{aligned} \tag{51}$$

with $\tilde{\phi}_n(x) \equiv 1_{\mathcal{C}_n}(x)\cdot(1+\epsilon)f_{\tilde{\mu}_n}(x)\log\frac{1}{(1-\epsilon)f_{\tilde{\mu}_n}(x)}$. The left term in (51) is upper bounded by $a_n(1+\epsilon)(H(v)+\log(1/a_n))$, which goes to zero with n from (a_n) being $o(1)$ and the fact that $v \in \mathcal{H}(\mathbb{X})$. For the right term in (51), (h_n) being $o(1)$ implies that x belongs to $\mathcal{B}_n$ eventually (in n) $\forall x \in A_v \setminus A_\mu$, then $\tilde{\phi}_n(x)$ tends to zero point-wise as n goes to infinity. On the other hand, for all $x \in \mathcal{C}_n$ (see (50)), we have that

$$\frac{1}{1/m+1} \leq \frac{\mu(A_n(x)\cap A_\mu)}{v(A_n(x)\cap A_\mu) + v(A_v\setminus A_\mu)} \leq \frac{\mu(A_n(x))}{v(A_n(x))} \leq \frac{\mu(A_n(x)\cap A_\mu)}{v(A_n(x)\cap A_\mu)} \leq M. \tag{52}$$

These inequalities derive from (48). Consequently for all $x \in \mathbb{X}$, if n sufficiently large such that $a_n < 0.5$, then

$$\begin{aligned}0 \leq \tilde{\phi}_n(x) &\leq (1+\epsilon)(a_n + (1-a_n)M)f_v(x)\log\frac{1}{(1-\epsilon)(a_n+(1-a_n)m/(m+1))} \\ &\leq (1+\epsilon)(1+M)f_v(x)\left[\log\frac{2(m+1)}{(1-\epsilon)} + \log\frac{1}{f_v(x)}\right].\end{aligned} \tag{53}$$

Hence from (50), $\tilde{\phi}_n(x)$ is bounded by a fix function that is $\ell_1(\mathbb{X})$ by the assumption that $v \in \mathcal{H}(\mathbb{X})$. Then by the dominated convergence theorems [43] and (51),

$$\lim_{n\to\infty}\sum_{\mathbb{X}}\phi_n^*(x) \leq \lim_{n\to\infty}\sum_{\mathbb{X}}\tilde{\phi}_n(x).$$

In summary, we have shown that for any arbitrary $(x_n) \in \mathcal{A}$ the sufficient conditions of Lemma 3 are satisfied, which proves the result in (25) reminding that $\mathbb{P}_\mu(\mathcal{A}) = 1$ from (45). □

Proof of Theorem 4. Let us first introduce the oracle probability

$$\mu_{\epsilon_n} \equiv \mu(\cdot|\Gamma_{\epsilon_n}) \in \mathcal{P}(\mathbb{X}). \tag{54}$$

Note that μ_{ϵ_n} is a random probability measure (function of the i.i.d sequence $X_1, \ldots, X_n$) as Γ_{ϵ_n} is a data-driven set, see (26). We will first show that:

$$\lim_{n\to\infty} H(\mu_{\epsilon_n}) = H(\mu) \text{ and } \lim_{n\to\infty} D(\mu_{\epsilon_n}||\mu) = 0, \ \mathbb{P}_\mu\text{-a.s.} \tag{55}$$

Under the assumption on (ϵ_n) of Theorem 4, $\lim_{n\to\infty} |\mu(\Gamma_{\epsilon_n}) - \hat{\mu}_n(\Gamma_{\epsilon_n})| = 0$, $\mathbb{P}_\mu$-a.s. (this result derives from the fact that $\lim_{n\to\infty} V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) = 0$, $\mathbb{P}_\mu$-a.s. , from (63)) In addition, since (ϵ_n) is $o(1)$ then $\lim_{n\to\infty} \hat{\mu}_n(\Gamma_{\epsilon_n}) = 1$, which implies that $\lim_{n\to\infty} \mu(\Gamma_{\epsilon_n}) = 1$ $\mathbb{P}_\mu$-a.s. From this $\mu_{\epsilon_n} \Rightarrow \mu$, $\mathbb{P}_\mu$-a.s. Let us consider a sequences (x_n) where $\lim_{n\to\infty} \mu(\Gamma_{\epsilon_n}) = 1$. Constrained to that

$$\limsup_{n\to\infty} \sup_{x\in A_\mu} \frac{f_{\mu_{\epsilon_n}}(x)}{f_\mu(x)} = \limsup_{n\to\infty} \frac{1}{\mu(\Gamma_{\epsilon_n})} < \infty. \tag{56}$$

Then there is $N > 0$ such that $\sup_{n>N} \sup_{x\in A_\mu} \frac{f_{\mu_{\epsilon_n}}(x)}{f_\mu(x)} < \infty$. Hence from Lemma 2, $\lim_{n\to\infty} D(\mu_{\epsilon_n}||\mu) = 0$ and $\lim_{n\to\infty} |H(\mu_{\epsilon_n}) - H(\mu)| = 0$. Finally, the set of sequences (x_n) where $\lim_{n\to\infty} \mu(\Gamma_{\epsilon_n}) = 1$ has probability one (with respect to $\mathbb{P}_\mu$), which proves (55).

For the rest of the proof, we concentrate on the analysis of $|H(\hat{\mu}_{n,\epsilon_n}) - H(\mu_{\epsilon_n})|$ that can be attributed to the estimation error aspect of the problem. It is worth noting that by construction $A_{\hat{\mu}_{n,\epsilon_n}} = A_{\mu_{\epsilon_n}} = \Gamma_{\epsilon_n}$, $\mathbb{P}_\mu$-a.s.. Consequently, we can use

$$H(\hat{\mu}_{n,\epsilon_n}) - H(\mu_{\epsilon_n}) = \sum_{x\in\Gamma_{\epsilon_n}} [\mu_{\epsilon_n}(\{x\}) - \hat{\mu}_{n,\epsilon_n}(\{x\})] \log \hat{\mu}_{n,\epsilon_n}(\{x\}) + D(\mu_{\epsilon_n}||\hat{\mu}_{n,\epsilon_n}). \tag{57}$$

The first term on the RHS of (57) is upper bounded by $\log 1/\mathbf{m}_{\hat{\mu}_n}^{\epsilon_n} \cdot V(\mu_{\epsilon_n}, \hat{\mu}_{n,\epsilon_n}) \leq \log 1/\epsilon_n \cdot V(\mu_{\epsilon_n}, \hat{\mu}_{n,\epsilon_n})$. Concerning the second term on the RHS of (57), it is possible to show (details presented in Appendix C) that

$$D(\mu_{\epsilon_n}||\hat{\mu}_{n,\epsilon_n}) \leq \frac{2\log\frac{e}{\epsilon_n}}{\mu(\Gamma_{\epsilon_n})} \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}), \tag{58}$$

where

$$V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) \equiv \sup_{A\in\sigma_{\epsilon_n}} |\mu(A) - \hat{\mu}_n(A)|. \tag{59}$$

In addition, it can be verified (details presented in Appendix D) that

$$V(\mu_{\epsilon_n}, \hat{\mu}_{n,\epsilon_n}) \leq K \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}), \tag{60}$$

for some universal constant $K > 0$. Therefore from (57), (58) and (60), there is $C > 0$ such that

$$|H(\hat{\mu}_{n,\epsilon_n}) - H(\mu_{\epsilon_n})| \leq \frac{C}{\mu(\Gamma_{\epsilon_n})} \log\frac{1}{\epsilon_n} \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}). \tag{61}$$

As mentioned before, $\mu(\Gamma_{\epsilon_n})$ goes to 1 almost surely, then we need to concentrate on the analysis of the asymptotic behavior of $\log 1/\epsilon_n \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n})$. From Hoeffding's inequality [28], we have that $\forall\delta > 0$

$$\mathbb{P}_\mu^n \left(\log 1/\epsilon_n \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) > \delta\right) \leq 2^{|\Gamma_{\epsilon_n}|+1} \cdot e^{-\frac{2n\delta^2}{(\log 1/\epsilon_n)^2}}, \tag{62}$$

considering that by construction $|\sigma_{\epsilon_n}| \leq 2^{|\Gamma_{\epsilon_n}|+1} \leq 2^{1/\epsilon_n+1}$. Assuming that (ϵ_n) is $O(n^{-\tau})$,

$$\ln \mathbb{P}_\mu^n \left(\log 1/\epsilon_n \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) > \delta \right) \leq (n^\tau + 1) \ln 2 - \frac{2n\delta^2}{\tau \log n}.$$

Therefore for all $\tau \in (0,1)$, $\delta > 0$ and any arbitrary $l \in (\tau, 1)$

$$\limsup_{n\to\infty} \frac{1}{n^l} \cdot \ln \mathbb{P}_\mu^n \left(\log 1/\epsilon_n \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) > \delta \right) < 0. \tag{63}$$

This last result is sufficient to show that $\sum_{n\geq 1} \mathbb{P}_\mu^n \left(\log 1/\epsilon_n \cdot V(\mu/\sigma_{\epsilon_n}, \hat{\mu}_n/\sigma_{\epsilon_n}) > \delta \right) < \infty$ that concludes the argument from the Borel-Cantelli Lemma. □

Proof of Theorem 5. We consider the expression

$$|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})| \leq |H(\mu) - H(\mu_{\epsilon_n})| + |H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})| \tag{64}$$

to analize the approximation error and the estimation error terms separately.

- Approximation Error Analysis

Note that $|H(\mu) - H(\mu_{\epsilon_n})|$ is a random object as μ_{ϵ_n} in (54) is a function of the data-dependent partition and, consequently, a function of $X_1, \ldots, X_n$. In the following, we consider the oracle set

$$\tilde{\Gamma}_{\epsilon_n} \equiv \{x \in \mathbb{X} : \mu(\{x\}) \geq \epsilon_n\}, \tag{65}$$

and the oracle conditional probability

$$\tilde{\mu}_{\epsilon_n} \equiv \mu(\cdot|\tilde{\Gamma}_{\epsilon_n}) \in \mathcal{P}.(\mathbb{X}). \tag{66}$$

Note that $\tilde{\Gamma}_{\epsilon_n}$ is a deterministic function of (ϵ_n) and so is the measure $\tilde{\mu}_{\epsilon_n}$ in (66). From definitions and triangular inequality:

$$\begin{aligned} |H(\mu) - H(\tilde{\mu}_{\epsilon_n})| \leq & \sum_{x\in\tilde{\Gamma}_{\epsilon_n}^c} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} + \log \frac{1}{\mu(\tilde{\Gamma}_{\epsilon_n})} \\ & + \left(\frac{1}{\mu(\tilde{\Gamma}_{\epsilon_n})} - 1 \right) \cdot \sum_{x\in\tilde{\Gamma}_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})}, \end{aligned} \tag{67}$$

and, similarly, the approximation error is bounded by

$$\begin{aligned} |H(\mu) - H(\mu_{\epsilon_n})| \leq & \sum_{x\in\Gamma_{\epsilon_n}^c} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} \\ & + \log \frac{1}{\mu(\Gamma_{\epsilon_n})} + \left(\frac{1}{\mu(\Gamma_{\epsilon_n})} - 1 \right) \cdot \sum_{x\in\Gamma_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})}. \end{aligned} \tag{68}$$

We denote the RHS of (67) and (68) by a_{ϵ_n} and $b_{\epsilon_n}(X_1, \ldots, X_n)$, respectively.
We can show that if (ϵ_n) is $O(n^{-\tau})$ and $\tau \in (0, 1/2)$, then

$$\limsup_{n\to\infty} b_{\epsilon_n}(X_1, \ldots, X_n) - a_{2\epsilon_n} \leq 0, \ \mathbb{P}_\mu\text{-a.s.}, \tag{69}$$

which from (68) implies that $|H(\mu) - H(\mu_{\epsilon_n})|$ is $O(a_{2\epsilon_n})$, $\mathbb{P}_\mu$-a.s. The proof of (69) is presented in Appendix E.

Then, we need to analyze the rate of convergence of the deterministic sequence $(a_{2\epsilon_n})$. Analyzing the RHS of (67), we recognize two independent terms: the partial entropy sum

$\sum_{x\in\tilde{\Gamma}^c_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})}$ and the rest that is bounded asymptotically by $\mu(\tilde{\Gamma}^c_{\epsilon_n})(1+H(\mu))$, using the fact that $\ln x \leq x-1$ for $x \geq 1$. Here is where the tail condition on μ plays a role. From the tail condition, we have that

$$\mu(\tilde{\Gamma}^c_{\epsilon_n}) \leq \mu\left(\left\{(k_o/\epsilon_n)^{1/p}+1, (k_o/\epsilon_n)^{1/p}+2, (k_o/\epsilon_n)^{1/p}+3, \ldots\right\}\right) = \sum_{x\geq(\frac{k_o}{\epsilon_n})^{1/p}+1} \mu(\{x\})$$
$$\leq k_1 \cdot \mathcal{S}_{(\frac{k_o}{\epsilon_n})^{1/p}+1}, \tag{70}$$

where $\mathcal{S}_{x_o} \equiv \sum_{x\geq x_o} x^{-p}$. Similarly as $\left\{0, 1, \ldots, (k_o/\epsilon_n)^{1/p}\right\} \subset \tilde{\Gamma}_{\epsilon_n}$, then

$$\sum_{x\in\tilde{\Gamma}^c_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} \leq \sum_{x\geq(\frac{k_o}{\epsilon_n})^{1/p}+1} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} \leq \sum_{x\geq(\frac{k_o}{\epsilon_n})^{1/p}+1} k_1 x^{-p} \cdot \log \frac{1}{k_0 x^{-p}}$$
$$\leq k_1 \log p \cdot \mathcal{R}_{(\frac{k_o}{\epsilon_n})^{1/p}+1} + k_1 \log 1/k_0 \cdot \mathcal{S}_{(\frac{k_o}{\epsilon_n})^{1/p}+1}, \tag{71}$$

where $\mathcal{R}_{x_o} \equiv \sum_{x\geq x_o} x^{-p} \log x$.

In Appendix F, it is shown that $\mathcal{S}_{x_o} \leq C_0 \cdot x_o^{1-p}$ and $\mathcal{R}_{x_o} \leq C_1 \cdot x_o^{1-p}$ for constants $C_1 > 0$ and $C_0 > 0$. Integrating these results in the RHS of (70) and (71) and considering that (ϵ_n) is $O(n^{-\tau})$, we have that both $\mu(\tilde{\Gamma}^c_{\epsilon_n})$ and $\sum_{x\in\tilde{\Gamma}^c_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})}$ are $O(n^{-\frac{\tau(p-1)}{p}})$. This implies that our oracle sequence (a_{ϵ_n}) is $O(n^{-\frac{\tau(p-1)}{p}})$.

In conclusion, if ϵ_n is $O(n^{-\tau})$ for $\tau \in (0, 1/2)$, it follows that

$$|H(\mu) - H(\mu_{\epsilon_n})| \text{ is } O(n^{-\frac{\tau(p-1)}{p}}), \ \mathbb{P}_\mu\text{-a.s.} \tag{72}$$

- Estimation Error Analysis

Let us consider $|H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})|$. From the bound in (61) and the fact that for any $\tau \in (0,1)$, $\lim_{n\to\infty} \mu(\Gamma_{\epsilon_n}) = 1$ $\mathbb{P}_\mu$-a.s. from (63), the problem reduces to analyze the rate of convergence of the following random object:

$$\rho_n(X_1, \ldots, X_n) \equiv \log \frac{1}{\epsilon_n} \cdot V(\mu/\sigma(\Gamma_{\epsilon_n}), \hat{\mu}_n/\sigma(\Gamma_{\epsilon_n})). \tag{73}$$

We will analize, instead, the oracle version of $\rho_n(X_1, \ldots, X_n)$ given by:

$$\xi_n(X_1, \ldots, X_n) \equiv \log \frac{1}{\epsilon_n} \cdot V(\mu/\sigma(\tilde{\Gamma}_{\epsilon_n/2}), \hat{\mu}_n/\sigma(\tilde{\Gamma}_{\epsilon_n/2})), \tag{74}$$

where $\tilde{\Gamma}_\epsilon \equiv \{x \in \mathbb{X} : \mu(\{x\}) \geq \epsilon\}$ is the oracle counterpart of Γ_ϵ in (26). To do so, we can show that if ϵ_n is $O(n^{-\tau})$ with $\tau \in (0, 1/2)$, then

$$\liminf_{n\to\infty} \xi_n(X_1, \ldots, X_n) - \rho_n(X_1, \ldots, X_n) \geq 0, \ \mathbb{P}_\mu\text{-a.s.}. \tag{75}$$

The proof of (75) is presented in Appendix G.

Moving to the almost sure rate of convergence of $\xi_n(X_1, \ldots, X_n)$, it is simple to show for our p-power dominating distribution that if (ϵ_n) is $O(n^{-\tau})$ and $\tau \in (0, p)$ then

$$\lim_{n\to\infty} \xi_n(X_1, \ldots, X_n) = 0 \ \mathbb{P}_\mu\text{-a.s.},$$

and, more specifically,

$$\xi_n(X_1, \ldots, X_n) \text{ is } o(n^{-q}) \text{ for all } q \in (0, (1-\tau/p)/2), \ \mathbb{P}_\mu\text{-a.s.}. \tag{76}$$

The argument is presented in Appendix H.

In conclusion, if ϵ_n is $O(n^{-\tau})$ for $\tau \in (0, 1/2)$, it follows that

$$|H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})| \text{ is } O(n^{-q}),\ \mathbb{P}_\mu\text{-a.s.}, \tag{77}$$

for all $q \in (0, (1-\tau/p)/2)$.

- Estimation vs. Approximation Errors

Coming back to (64) and using (72) and (77), the analysis reduces to finding the solution τ^* in $(0, 1/2)$ that offers the best trade-off between the estimation and approximation error rate:

$$\tau^* \equiv \arg \max_{\tau(0,1/2)} \min \left\{ \frac{(1-\tau/p)}{2}, \frac{\tau(p-1)}{p} \right\}. \tag{78}$$

It is simple to verify that $\tau^* = 1/2$. Then by considering τ arbitrary close to the admissible limit $1/2$, we can achieve a rate of convergence for $|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})|$ that is arbitrary close to $O(n^{-\frac{1}{2}(1-1/p)})$, $\mathbb{P}$-a.s.

More formally, for any $l \in (0, \frac{1}{2}(1-1/p))$ we can take $\tau \in (\frac{l}{(1-1/p)}, \frac{1}{2})$ where $|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})|$ is $o(n^{-l})$, $\mathbb{P}_\mu$-a.s., from (72) and (77).

Finally, a simple corollary of this analysis is to consider $\tau(p) = \frac{1}{2+1/p} < 1/2$ where:

$$|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})| \text{ is } O(n^{-\frac{1-1/p}{2+1/p}}),\ \mathbb{P}_\mu\text{-a.s.}, \tag{79}$$

which concludes the argument. □

Proof of Theorem 6. The argument follows the proof of Theorem 5. In particular, we use the estimation-approximation error bound:

$$|H(\mu) - H(\hat{\mu}_{n,\epsilon_n})| \leq |H(\mu) - H(\mu_{\epsilon_n})| + |H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})|, \tag{80}$$

and the following two results derived in the proof of Theorem 5: If (ϵ_n) is $O(n^{-\tau})$ with $\tau \in (0, 1/2)$ then (for the approximation error)

$$|H(\mu) - H(\mu_{\epsilon_n})| \text{ is } O(a_{2\epsilon_n})\ \mathbb{P}_\mu\text{-a.s.}, \tag{81}$$

with $a_{\epsilon_n} = \sum_{x \in \tilde{\Gamma}^c_{\epsilon_n}} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} + \mu(\tilde{\Gamma}^c_{\epsilon_n})(1 + H(\mu))$, while (for the estimation error)

$$|H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})| \text{ is } O(\xi_n(X_1, \ldots, X_n))\ \mathbb{P}_\mu\text{-a.s.}, \tag{82}$$

with $\xi_n(X_1, \ldots, X_n) = \log \frac{1}{\epsilon_n} \cdot V(\mu/\sigma(\tilde{\Gamma}_{\epsilon_n/2}), \hat{\mu}_n/\sigma(\tilde{\Gamma}_{\epsilon_n/2}))$.

For the estimation error, we need to bound the rate of convergence of $\xi_n(X_1, \ldots, X_n)$ to zero almost surely. We first note that $\{1, \ldots, x_o(\epsilon_n)\} = \tilde{\Gamma}_{\epsilon_n}$ with $x_o(\epsilon_n) = \lfloor 1/\alpha \ln(k_0/\epsilon_n) \rfloor$. Then from Hoeffding's inequality we have that

$$\begin{aligned} \mathbb{P}^n_\mu(\{\xi_n(X_1, \ldots, X_n) > \delta\}) &\leq 2^{|(\tilde{\Gamma}_{\epsilon_n/2})|} \cdot e^{-2n\frac{\delta^2}{\log(1/\epsilon_n)^2}} \\ &\leq 2^{1/\alpha \ln(2k_0/\epsilon_n)+1} \cdot e^{-2n\frac{\delta^2}{\log(1/\epsilon_n)^2}}. \end{aligned} \tag{83}$$

Considering $\epsilon_n = O(n^{-\tau})$, an arbitrary sequence (δ_n) being $o(1)$ and $l > 0$, it follows from (83) that

$$\frac{1}{n^l} \cdot \ln \mathbb{P}^n_\mu(\{\xi_n(X_1, \ldots, X_n) > \delta_n\}) \leq \frac{1}{n^l} \ln(2) \left[1/\alpha \ln(2k_0/\epsilon_n) + 1\right] - n^{1-l} \frac{\delta_n^2}{\log(1/\epsilon_n)^2}. \tag{84}$$

We note that the first term in the RHS of (84) is $O(\frac{1}{n^l}\log n)$ and goes to zero for all $l > 0$, while the second term is $O(n^{1-l}\frac{\delta_n^2}{\log n^2})$. If we consider $\delta_n = O(n^{-q})$, this second term is $O(n^{1-2q-l}\cdot\frac{1}{\log n^2})$. Therefore, for any $q \in (0,1/2)$ we can take an arbitrary $l \in (0, 1-2q]$ such that $\mathbb{P}_\mu^n(\{\xi_n(X_1,\ldots,X_n) > \delta_n\})$ is $O(e^{-n^l})$ from (84). This result implies, from the Borel-Cantelli Lemma, that $\xi_n(X_1,\ldots,X_n)$ is $o(\delta_n)$, $\mathbb{P}_\mu$-a.s, which in summary shows that $|H(\mu_{\epsilon_n}) - H(\hat{\mu}_{n,\epsilon_n})|$ is $O(n^{-q})$ for all $q \in (0,1/2)$.

For the approximation error, it is simple to verify that:

$$\mu(\tilde{\Gamma}_{\epsilon_n}^c) \leq k_1 \cdot \sum_{x \geq x_o(\epsilon_n)+1} e^{-\alpha x} = k_1 \cdot \tilde{\mathcal{S}}_{x_o(\epsilon_n)+1} \tag{85}$$

and

$$\sum_{x\in\tilde{\Gamma}_{\epsilon_n}^c} \mu(\{x\}) \log \frac{1}{\mu(\{x\})} \leq \sum_{x \geq x_o(\epsilon_n)+1} k_1 e^{-\alpha x} \log \frac{1}{k_0 e^{-\alpha x}} = k_1 \log\frac{1}{k_0}\cdot\tilde{\mathcal{S}}_{x_o(\epsilon_n)+1} + \alpha \log e \cdot k_1 \cdot \tilde{\mathcal{R}}_{x_o(\epsilon_n)+1}, \tag{86}$$

where $\tilde{\mathcal{S}}_{x_o} \equiv \sum_{x\geq x_o} e^{-\alpha x}$ and $\tilde{\mathcal{R}}_{x_o} \equiv \sum_{x\geq x_o} x\cdot e^{-\alpha x}$. At this point, it is not difficult to show that $\tilde{\mathcal{S}}_{x_o} \leq M_1 e^{-\alpha x_o}$ and $\tilde{\mathcal{R}}_{x_o} \leq M_2 e^{-\alpha x_o}\cdot x_o$ for some constants $M_1 > 0$ and $M_2 > 0$. Integrating these partial steps, we have that

$$a_{\epsilon_n} \leq k_1(1 + H(\mu) + \log\frac{1}{k_0})\cdot\tilde{\mathcal{S}}_{x_o(\epsilon_n)+1} + \alpha\log e\cdot k_1\cdot\tilde{\mathcal{R}}_{x_o(\epsilon_n)+1} \leq O_1\cdot\epsilon_n + O_2\cdot\epsilon_n\log\frac{1}{\epsilon_n} \tag{87}$$

for some constant $O_1 > 0$ and $O_2 > 0$. The last step is from the evaluation of $x_o(\epsilon_n) = \lfloor 1/\alpha \ln(k_0/\epsilon_n)\rfloor$. Therefore from (81) and (87), it follows that $|H(\mu) - H(\mu_{\epsilon_n})|$ is $O(n^{-\tau}\log n)$ $\mathbb{P}_\mu$-a.s. for all $\tau \in (0,1/2)$.

The argument concludes by integrating in (80) the almost sure convergence results obtained for the estimation and approximation errors. □

Proof of Theorem 7. Let us define the event

$$\mathcal{B}_n^\epsilon = \left\{x^n \in \mathbb{X}^n : \Gamma_\epsilon(x^n) = A_\mu\right\}, \tag{88}$$

that represents the detection of the support of μ from the data for a given $\epsilon > 0$ in (26). Note that the dependency on the data for Γ_ϵ is made explicit in this notation. In addition, let us consider the deviation event

$$\mathcal{A}_n^\epsilon(\mu) = \{x^n \in \mathbb{X}^n : V(\mu,\hat{\mu}_n) > \epsilon\}. \tag{89}$$

By the hypothesis that $|A_\mu| < \infty$, then $\mathbf{m}_\mu = \min_{x\in A_\mu} f_\mu(x) > 0$. Therefore if $x^n \in (\mathcal{A}_n^{\mathbf{m}_\mu/2}(\mu))^c$ then $\hat{\mu}_n(\{x\}) \geq \mathbf{m}_\mu/2$ for all $x \in A_\mu$, which implies that $(\mathcal{B}_n^\epsilon)^c \subset \mathcal{A}_n^{\mathbf{m}_\mu/2}(\mu)$ as long as $0 < \epsilon \leq \mathbf{m}_\mu/2$. Using the hypothesis that $\epsilon_n \to 0$, there is $N > 0$ such that for all $n \geq N$ $(\mathcal{B}_n^{\epsilon_n})^c \subset \mathcal{A}_n^{\mathbf{m}_\mu/2}(\mu)$ and, consequently,

$$\mathbb{P}_\mu^n((\mathcal{B}_n^{\epsilon_n})^c) \leq \mathbb{P}_\mu^n(\mathcal{A}_n^{\mathbf{m}_\mu/2}(\mu)) \leq 2^{k+1}\cdot e^{-\frac{n\mathbf{m}_\mu^2}{4}}, \tag{90}$$

the last from Hoeffding's inequality considering $k = |A_\mu| < \infty$.

If we consider the events:

$$\mathcal{C}_n^\epsilon(\mu) = \{x^n \in \mathbb{X}^n : |H(\hat{\mu}_{n,\epsilon_n}) - H(\mu)| > \epsilon\} \text{ and} \tag{91}$$

$$\mathcal{D}_n^\epsilon(\mu) = \{x^n \in \mathbb{X}^n : |H(\hat{\mu}_n) - H(\mu)| > \epsilon\} \tag{92}$$

and we use the fact that by definition $\hat{\mu}_{n,\epsilon_n} = \hat{\mu}_n$ conditioning on $\mathcal{B}_n^{\epsilon_n}$, it follows that $\mathcal{C}_n^{\epsilon}(\mu) \cap \mathcal{B}_n^{\epsilon_n} \subset \mathcal{D}_n^{\epsilon}(\mu)$. Then, for all $\epsilon > 0$ and $n \geq N$

$$\begin{aligned}\mathbb{P}_\mu^n(\mathcal{C}_n^\epsilon(\mu)) &\leq \mathbb{P}_\mu^n(\mathcal{C}_n^\epsilon(\mu) \cap \mathcal{B}_n^{\epsilon_n}) + \mathbb{P}_\mu^n((\mathcal{B}_n^{\epsilon_n})^c) \\ &\leq \mathbb{P}_\mu^n(\mathcal{D}_n^\epsilon(\mu)) + \mathbb{P}_\mu^n((\mathcal{B}_n^{\epsilon_n})^c) \\ &\leq 2^{k+1}\left[e^{-\frac{2n\epsilon^2}{(\mathbf{M}_\mu + \frac{\log e}{\mathbf{m}_\mu})^2}} + e^{-\frac{n\mathbf{m}_\mu^2}{4}}\right], \end{aligned} \tag{93}$$

the last inequality from Theorem 1 and (90). □

Funding: The work is supported by funding from FONDECYT Grant 1170854, CONICYT-Chile and the Advanced Center for Electrical and Electronic Engineering (AC3E), Basal Project FB0008.

Acknowledgments: The author is grateful to Patricio Parada for his insights and stimulating discussion in the initial stage of this work. The author thanks the anonymous reviewers for their valuable comments and suggestions, and his colleagues Claudio Estevez, Rene Mendez and Ruben Claveria for proofreading this material.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Minimax risk for Finite Entropy Distributions in ∞-Alphabets

Proposition A1. $R_n^* = \infty$.

For the proof, we use the following lemma that follows from [26] (Theorem 1).

Lemma A1. *Let us fix two arbitrary real numbers $\delta > 0$ and $\epsilon > 0$. Then there are P, Q two finite supported distributions on $\mathbb{H}(\mathbb{X})$ that satisfy that $D(P||Q) < \epsilon$ while $H(Q) - H(P) > \delta$.*

The proof of Lemma A1 derives from the same construction presented in the proof of [26] (Theorem 1), i.e., $P = (p_1, \ldots, p_L)$ and a modification of it $Q_M = (p_1 \cdot (1 - 1/\sqrt{M}), p_2 + p_1/M\sqrt{M}, \ldots, p_L + p_1/M\sqrt{M}, p_1/M\sqrt{M}, \ldots, p_1/M\sqrt{M})$ both distribution of finite support and consequently in $\mathbb{H}(\mathbb{X})$. It is simple to verify that as M goes to infinity $D(P||Q_M) \longrightarrow 0$ while $H(Q_M) - H(P) \longrightarrow \infty$.

Proof. For any pair of distribution P, Q in $\mathbb{H}(\mathbb{X})$, Le Cam's two point method [53] shows that:

$$R_n^* \geq \frac{1}{4}\left(H(Q) - H(P)\right)^2 \exp^{-nD(P||Q)}. \tag{A1}$$

Adopting Lemma A1 and Equation (A1), for any n and any arbitrary $\epsilon > 0$ and $\delta > 0$, we have that $R_n^* > \delta^2 \exp^{-n\epsilon}/4$. Then exploiting the discontinuity of the entropy in infinite alphabets, we can fix ϵ and make δ arbitrar large. □

Appendix B. Proposition A2

Proposition A2. *Under the assumptions of Theorem 3:*

$$\lim_{n\to\infty} \sup_{x \in A_{\tilde{\mu}_n}} \left| \frac{d\tilde{\mu}_n}{d\mu_n^*}(x) - 1 \right| = 0, \ \mathbb{P}_\mu\text{-a.s.} \tag{A2}$$

Proof. First note that $A_{\tilde{\mu}_n} = A_{\mu_n^*}$, then $\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)$ is finite and $\forall x \in A_{\tilde{\mu}_n}$

$$\frac{d\tilde{\mu}_n}{d\mu_n^*}(x) = \frac{(1-a_n)\cdot\mu(A_n(x)) + a_n v(A_n(x))}{(1-a_n)\cdot\hat{\mu}_n(A_n(x)) + a_n v(A_n(x))}. \tag{A3}$$

Then by construction,

$$\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)-1\right| \leq \sup_{A\in\pi_n}\frac{|\hat{\mu}_n(A)-\mu(A)|}{a_n\cdot h_n}. \tag{A4}$$

From Hoeffding's inequality, we have that $\forall\epsilon>0$

$$\mathbb{P}_\mu^n\left(\sup_{A\in\pi_n}|\hat{\mu}_n(A)-\mu(A)|>\epsilon\right)\leq 2\cdot|\pi_n|\cdot\exp^{-2n\epsilon^2}. \tag{A5}$$

By condition ii), given that $(1/a_nh_n)$ is $o(n^\tau)$ for some $\tau\in(0,1/2)$, then there exists $\tau_o\in(0,1)$ such that

$$\lim_{n\to\infty}\frac{1}{n^{\tau_o}}\ln\mathbb{P}_\mu^n\left(\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)-1\right|>\epsilon\right)\leq\lim_{n\to\infty}\frac{1}{n^{\tau_o}}\ln(2\,|\pi_n|)-2\cdot(n^{\frac{1-\tau_o}{2}}a_nh_n\epsilon)^2=-\infty.$$

This implies that $\mathbb{P}_\mu^n\left(\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)-1\right|>\epsilon\right)$ is eventually dominated by a constant time $(e^{-n^{\tau_o}})_{n\geq1}$, which from the Borel-Cantelli Lemma [43] implies that

$$\lim_{n\to\infty}\sup_{x\in A_{\tilde{\mu}_n}}\left|\frac{d\tilde{\mu}_n}{d\mu_n^*}(x)-1\right|=0,\ \mathbb{P}_\mu\text{-a.s.}. \tag{A6}$$

□

Appendix C. Proposition A3

Proposition A3.

$$D(\mu_{\epsilon_n}||\hat{\mu}_{n,\epsilon_n})\leq\frac{2\log\frac{e}{\epsilon_n}}{\mu(\Gamma_{\epsilon_n})}\cdot V(\mu/\sigma_{\epsilon_n},\hat{\mu}_n/\sigma_{\epsilon_n})$$

Proof. From definition,

$$D(\mu_{\epsilon_n}||\hat{\mu}_{n,\epsilon_n})=\frac{1}{\mu(\Gamma_{\epsilon_n})}\sum_{x\in\Gamma_{\epsilon_n}}f_\mu(x)\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)}+\log\frac{\hat{\mu}_n(\Gamma_{\epsilon_n})}{\mu(\Gamma_{\epsilon_n})}. \tag{A7}$$

For the right term in the RHS of (A7):

$$\log\frac{\hat{\mu}_n(\Gamma_{\epsilon_n})}{\mu(\Gamma_{\epsilon_n})}\leq\frac{\log(e)}{\mu(\Gamma_{\epsilon_n})}\left|\hat{\mu}_n(\Gamma_{\epsilon_n})-\mu(\Gamma_{\epsilon_n})\right|. \tag{A8}$$

For the left term in the RHS of (A7):

$$\begin{aligned}
\left|\sum_{x\in\Gamma_{\epsilon_n}} f_\mu(x)\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)}\right| &= \left|\sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)\le f_{\hat{\mu}_n}(x)}} f_\mu(x)\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)} + \sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)>f_{\hat{\mu}_n}(x)\ge\epsilon_n}} f_\mu(x)\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)}\right| \\
&\le \sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)\le f_{\hat{\mu}_n}(x)}} f_\mu(x)\log\frac{f_{\hat{\mu}_n}(x)}{f_\mu(x)} + \sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)>f_{\hat{\mu}_n}(x)\ge\epsilon_n}} f_{\hat{\mu}_n}(x)\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)} \\
&\quad + \sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)>f_{\hat{\mu}_n}(x)\ge\epsilon_n}} (f_\mu(x)-f_{\hat{\mu}_n}(x))\cdot\log\frac{f_\mu(x)}{f_{\hat{\mu}_n}(x)} \qquad \text{(A9)} \\
&\le \log e\left[\sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)\le f_{\hat{\mu}_n}(x)}} (f_{\hat{\mu}_n}(x)-f_\mu(x)) + \sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)>f_{\hat{\mu}_n}(x)}} (f_\mu(x)-f_{\hat{\mu}_n}(x))\right] \\
&\quad + \log\tfrac{1}{\epsilon_n}\cdot\sum_{\substack{x\in\Gamma_{\epsilon_n}\\ f_\mu(x)>f_{\hat{\mu}_n}(x)}} (f_\mu(x)-f_{\hat{\mu}_n}(x)) \qquad \text{(A10)} \\
&\le (\log e+\log\tfrac{1}{\epsilon_n})\cdot\sum_{x\in\Gamma_{\epsilon_n}}\left|f_\mu(x)-f_{\hat{\mu}_n}(x)\right|. \qquad \text{(A11)}
\end{aligned}$$

The first inequality in (A9) is by triangular inequality, the second in (A10) is from the fact that $\ln x \le x-1$ for $x>0$. Finally, from definition of the total variational distance over σ_{ϵ_n} in (59) we have that

$$2\cdot V(\mu/\sigma_{\epsilon_n},\hat{\mu}_n/\sigma_{\epsilon_n}) = \sum_{x\in\Gamma_{\epsilon_n}}\left|f_\mu(x)-f_{\hat{\mu}_n}(x)\right| + \left|\hat{\mu}_n(\Gamma_{\epsilon_n})-\mu(\Gamma_{\epsilon_n})\right|, \qquad \text{(A12)}$$

which concludes the argument from (A7)–(A9). □

Appendix D. Proposition A4

Proposition A4. *Considering that $(k_n)\to\infty$, there exists $K>0$ and $N>0$ such that $\forall n\ge N$,*

$$V(\tilde{\mu}_{k_n},\hat{\mu}^*_{k_n,n}) \le K\cdot V(\mu/\sigma_{k_n},\hat{\mu}_n/\sigma_{k_n}). \qquad \text{(A13)}$$

Proof.

$$\begin{aligned}
V(\tilde{\mu}_{k_n},\hat{\mu}^*_{k_n,n}) &= \frac{1}{2}\sum_{x\in A_\mu\cap\Gamma_{k_n}}\left|\frac{\mu\{x\}}{\mu(\Gamma_{k_n})}-\frac{\hat{\mu}_n\{x\}}{\hat{\mu}_n(\Gamma_{k_n})}\right| \\
&\le \frac{1}{2\mu(\Gamma_{k_n})}\left[\sum_{x\in A_\mu\cap\Gamma_{k_n}}|\hat{\mu}_n\{x\}-\mu\{x\}| + \sum_{x\in A_\mu\cap\Gamma_{k_n}}\hat{\mu}_n\{x\}\left|\frac{\mu(\Gamma_{k_n})}{\hat{\mu}_n(\Gamma_{k_n})}-1\right|\right] \\
&= \frac{1}{2\mu(\Gamma_{k_n})}\left[2\cdot V(\mu/\sigma_{k_n},\hat{\mu}_n/\sigma_{k_n}) + \left|\mu(\Gamma_{k_n})-\hat{\mu}_n(\Gamma_{k_n})\right|\right] \\
&\le \frac{3\cdot V(\mu/\sigma_{k_n},\hat{\mu}_n/\sigma_{k_n})}{2\mu(\Gamma_{k_n})}. \qquad \text{(A14)}
\end{aligned}$$

By the hypothesis $\mu(\Gamma_{k_n})\to 1$, which concludes the proof. □

Appendix E. Proposition A5

Proposition A5. *If ϵ_n is $O(n^{-\tau})$ with $\tau\in(0,1/2)$, then*

$$\limsup_{n\to\infty} b_{\epsilon_n}(X_1,\ldots,X_n) - a_{2\epsilon_n} \le 0,\ \mathbb{P}_\mu - a.s..$$

Proof. Let us define the set

$$\mathcal{B}_n = \left\{(x_1,\ldots,x_n) : \tilde{\Gamma}_{2\epsilon_n} \subset \Gamma_{\epsilon_n}\right\} \subset \mathbb{X}^n.$$

From definition every sequence $(x_1,\ldots,x_n) \in \mathcal{B}_n$ is such that $b_{\epsilon_n}(x_1,\ldots,x_n) \leq a_{2\epsilon_n}$ and, consequently, we just need to prove that $\mathbb{P}_\mu(\liminf_{n\to\infty} \mathcal{B}_n) = \mathbb{P}_\mu(\cup_{n\geq 1} \cap_{k\geq n} \mathcal{B}_k) = 1$ [42]. Furthermore, if $\sup_{x\in\tilde{\Gamma}_{2\epsilon_n}} |\hat{\mu}_n(\{x\}) - \mu(\{x\})| \leq \epsilon_n$, then by definition of $\tilde{\Gamma}_{2\epsilon_n}$ in (65), we have that $\hat{\mu}_n(\{x\}) \geq \epsilon_n$ for all $x \in \Gamma_{2\epsilon_n}$ (i.e., $\tilde{\Gamma}_{2\epsilon_n} \subset \Gamma_{\epsilon_n}$). From this

$$\mathbb{P}_\mu^n(\mathcal{B}_n^c) \leq \mathbb{P}_\mu^n\left(\sup_{x\in\tilde{\Gamma}_{2\epsilon_n}} |\hat{\mu}_n(\{x\}) - \mu(\{x\})| > \epsilon_n\right) \leq \left|\tilde{\Gamma}_{2\epsilon_n}\right| \cdot e^{-2n\epsilon_n^2} \leq \frac{1}{2\epsilon_n} \cdot e^{-2n\epsilon_n^2}, \tag{A15}$$

from the Hoeffding's inequality [28,52], the union bound and the fact that by construction $\left|\tilde{\Gamma}_{2\epsilon_n}\right| \leq \frac{1}{2\epsilon_n}$. If we consider $\epsilon_n = O(n^{-\tau})$ and $l > 0$, we have that:

$$\frac{1}{n^l} \cdot \ln \mathbb{P}_\mu^n(\mathcal{B}_n^c) \leq \frac{1}{n^l} \ln(1/2 \cdot n^\tau) - 2n^{1-2\tau-l}. \tag{A16}$$

From (A16) for any $\tau \in (0,1/2)$ there is $l \in (0, 1-2\tau]$ such that $\mathbb{P}_\mu^n(\mathcal{B}_n^c)$ is bounded by a term $O(e^{-n^l})$. This implies that $\sum_{n\geq 1} \mathbb{P}_\mu^n(\mathcal{B}_n^c) < \infty$, that suffices to show that $\mathbb{P}_\mu(\cup_{n\geq 1} \cap_{k\geq n} \mathcal{B}_k) = 1$. □

Appendix F. Auxiliary Results for Theorem 5

Let us first consider the series

$$\begin{aligned}\mathcal{S}_{x_o} = \sum_{x\geq x_o} x^{-p} &= x_o^{-p} \cdot \left(1 + \left(\frac{x_o}{x_o+1}\right)^p + \left(\frac{x_o}{x_o+2}\right)^p + \ldots\right) \\ &= x_o^{-p} \cdot \left(\tilde{\mathcal{S}}_{x_o,0} + \tilde{\mathcal{S}}_{x_o,1} + \ldots + \tilde{\mathcal{S}}_{x_o,x_o-1}\right),\end{aligned} \tag{A17}$$

where $\tilde{\mathcal{S}}_{x_o,j} \equiv \sum_{k=0}^\infty \left(\frac{k\cdot x_o + j}{x_o}\right)^{-p}$ for all $j \in \{0,\ldots,x_o-1\}$. It is simple to verify that for all $j \in \{0,\ldots,x_o-1\}$, $\tilde{\mathcal{S}}_{x_o,j} \leq \tilde{\mathcal{S}}_{x_o,0} = \sum_{k\geq 0} k^{-p} < \infty$ given that by hypothesis $p > 1$. Consequently, $\mathcal{S}_{x_o} \leq x_o^{1-p} \cdot \sum_{k\geq 0} k^{-p}$.

Similarly, for the second series we have that:

$$\begin{aligned}\mathcal{R}_{x_o} = \sum_{x\geq x_o} x^{-p}\log x &= x_o^{-p} \cdot \left(\log(x_o) + \left(\frac{x_o}{x_o+1}\right)\log(x_o+1) + \left(\frac{x_o}{x_o+2}\right)\log(x_o+2) + \ldots\right) \\ &= x_o^{-p} \cdot \left(\tilde{\mathcal{R}}_{x_o,0} + \tilde{\mathcal{R}}_{x_o,2} + \ldots + \tilde{\mathcal{R}}_{x_o,x_o-1}\right),\end{aligned} \tag{A18}$$

where $\tilde{\mathcal{R}}_{x_o,j} \equiv \sum_{k=1}^\infty \left(\frac{k\cdot x_o + j}{x_o}\right)^{-p} \cdot \log(kx_o + j)$ for all $j \in \{0,\ldots,x_o-1\}$. Note again that $\tilde{\mathcal{R}}_{x_o,j} \leq \tilde{\mathcal{R}}_{x_o,0} < \infty$ for all $j \in \{0,\ldots,x_o-1\}$, and, consequently, $\mathcal{R}_{x_o} \leq x_o^{1-p} \cdot \sum_{k\geq 1} k^{-p}\log k$ from (A18).

Appendix G. Proposition A6

Proposition A6. *If ϵ_n is $O(n^{-\tau})$ with $\tau \in (0,1/2)$, then*

$$\liminf_{n\to\infty} \xi_n(X_1,\ldots,X_n) - \rho_n(X_1,\ldots,X_n) \geq 0, \ \mathbb{P}_\mu - a.s..$$

Proof. By definition if $\sigma(\Gamma_{\epsilon_n}) \subset \sigma(\tilde{\Gamma}_{\epsilon_n/2})$ then $\xi_n(X_1,\ldots,X_n) \geq \rho_n(X_1,\ldots,X_n)$. Consequently, if we define the set:

$$\mathcal{B}_n = \left\{(x_1,\ldots,x_n) : \sigma(\Gamma_{\epsilon_n}) \subset \sigma(\tilde{\Gamma}_{\epsilon_n/2})\right\}, \tag{A19}$$

then the proof reduced to verify that $\mathbb{P}_\mu(\liminf_{n\to\infty} \mathcal{B}_n) = \mathbb{P}_\mu(\cup_{n\geq 1} \cap_{k\geq n} \mathcal{B}_k) = 1$.

On the other hand, if $\sup_{x\in\Gamma_{\epsilon_n}} |\hat{\mu}_n(\{x\}) - \mu(\{x\})| \leq \epsilon_n/2$ then by definition of Γ_ϵ, for all $x \in \Gamma_{\epsilon_n}$ $\mu(\{x\}) \geq \epsilon_n/2$, i.e., $\Gamma_{\epsilon_n} \subset \tilde{\Gamma}_{\epsilon_n/2}$. In other words,

$$\mathcal{C}_n = \left\{ (x_1, \ldots, x_n) : \sup_{x\in\Gamma_{\epsilon_n}} |\hat{\mu}_n(\{x\}) - \mu(\{x\})| \leq \epsilon_n/2 \right\} \subset \mathcal{B}_n. \tag{A20}$$

Finally,

$$\mathbb{P}_\mu^n(\mathcal{C}_n^c) = \mathbb{P}_\mu^n \left(\sup_{x\in\Gamma_{\epsilon_n}} |\hat{\mu}_n(\{x\}) - \mu(\{x\})| > \epsilon_n/2 \right) \leq |\Gamma_{\epsilon_n}| \cdot e^{-n\epsilon^2/2} \leq \frac{1}{\epsilon_n} \cdot e^{-n\epsilon^2/2}. \tag{A21}$$

In this context, if we consider $\epsilon_n = O(n^{-\tau})$ and $l > 0$, then we have that:

$$\frac{1}{n^l} \cdot \ln \mathbb{P}_\mu^n(\mathcal{C}_n^c) \leq \tau \cdot \frac{\ln n}{n^l} - \frac{n^{1-2\tau-l}}{2}. \tag{A22}$$

Therefore, we have that for any $\tau \in (0, 1/2)$ we can take $l \in (0, 1-2\tau]$ such that $\mathbb{P}_\mu^n(\mathcal{C}_n^c)$ is bounded by a term $O(e^{-n^l})$. Then, the Borel Cantelli Lemma tells us that $\mathbb{P}_\mu(\cup_{n\geq 1} \cap_{k\geq n} \mathcal{C}_k) = 1$, which concludes the proof from (A20). □

Appendix H. Proposition A7

Proposition A7. *For the p-power tail dominating distribution stated in Theorem 5, if (ϵ_n) is $O(n^{-\tau})$ with $\tau \in (0, p)$ then $\xi_n(X_1, \ldots, X_n)$ is $o(n^{-q})$ for all $q \in (0, (1-\tau/p)/2)$, $\mathbb{P}_\mu$-a.s.*

Proof. From the Hoeffding's inequality we have that

$$\begin{aligned} \mathbb{P}_\mu^n \left(\{x_1, \ldots, x_n : \xi_n(x_1, \ldots, x_n) > \delta\} \right) &\leq \left| \sigma(\tilde{\Gamma}_{\epsilon_n/2}) \right| \cdot e^{-2n\frac{\delta^2}{\log(1/\epsilon_n)^2}} \\ &\leq 2^{(\frac{2k_0}{\epsilon_n})^{1/p}+1} \cdot e^{-2n\frac{\delta^2}{\log(1/\epsilon_n)^2}}, \end{aligned} \tag{A23}$$

the second inequality using that $\tilde{\Gamma}_\epsilon \leq (\frac{k_0}{\epsilon})^{1/p} + 1$ from the definition of $\tilde{\Gamma}_\epsilon$ in (65) and the tail bounded assumption on μ. If we consider $\epsilon_n = O(n^{-\tau})$ and $l > 0$, then we have that:

$$\frac{1}{n^l} \cdot \ln \mathbb{P}_\mu^n \left(\{x_1, \ldots, x_n : \xi_n(x_1, \ldots, x_n) > \delta\} \right) \leq \ln 2 \cdot (Cn^{\tau/p-l} + n^{-l}) - \frac{2\delta^2}{\tau^2} \cdot \frac{n^{1-l}}{\log n^2} \tag{A24}$$

for some constant $C > 0$. Then in order to obtain that $\xi_n(X_1, \ldots, X_n)$ converges almost surely to zero from (A24), it is sufficient that $l > 0$, $l < 1$, and $l > \tau/p$. This implies that if $\tau < p$, there is $l \in (\tau/p, 1)$ such that such that $\mathbb{P}_\mu^n(\xi_n(x_1, \ldots, x_n) > \delta)$ is bounded by a term $O(e^{-n^l})$ and, consequently, $\lim_{n\to\infty} \xi_n(X_1, \ldots, X_n) = 0$, $\mathbb{P}_\mu$-a.s. (this by using the same steps used in Appendix G).

Moving to the rate of convergence of $\xi_n(X_1, \ldots, X_n)$ (assuming that $\tau < p$), let us consider $\delta_n = n^{-q}$ for some $q \geq 0$. From (A24):

$$\frac{1}{n^l} \cdot \ln \mathbb{P}_\mu^n \left(\{x_1, \ldots, x_n : \xi_n(x_1, \ldots, x_n) > \delta_n\} \right) \leq \ln 2 \cdot (Cn^{\tau/p-l} + n^{-l}) - \frac{2\delta^2}{\tau^2} \cdot \frac{n^{1-2q-l}}{\log n^2}. \tag{A25}$$

To make $\xi_n(X_1, \ldots, X_n)$ being $o(n^{-q})$ $\mathbb{P}$-a.s., a sufficient condition is that $l > 0$, $l > \tau/p$, and $l < 1-2q$. Therefore (considering that $\tau < p$), the admissibility condition on the existence of a exponential rate of convergence $O(e^{-n^l})$ for $l > 0$ for the deviation event $\{x_1, \ldots, x_n : \xi_n(x_1, \ldots, x_n) > \delta_n\}$ is that $\tau/p < 1-2q$, which is equivalent to $0 < q < \frac{1-\tau/p}{2}$. □

References

1. Beirlant, J.; Dudewicz, E.; Györfi, L.; van der Meulen, E.C. Nonparametric entropy estimation: An Overview. *Int. Math. Stat. Sci.* **1997**, *6*, 17–39.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA. 2006.
3. Kullback, S. *Information Theory and Statistics*; Wiley: New York, NY, USA, 1959.
4. Principe, J. *Information Theoretic Learning: Renyi Entropy and Kernel Perspective*; Springer: New York, NY, USA, 2010.
5. Fisher, J.W., III; Wainwright, M.; Sudderth, E.; Willsky, A.S. Statistical and information-theoretic methods for self-organization and fusion of multimodal, networked sensors. *Int. J. High Perform. Comput. Appl.* **2002**, *16*, 337–353. [CrossRef]
6. Liu, J.; Moulin, P. Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Trans. Image Process.* **2001**, *10*, 1647–1658. [PubMed]
7. Thévenaz, P.; Unser, M. Optimization of mutual information for multiresolution image registration. *IEEE Trans. Image Process.* **2000**, *9*, 2083–2099. [PubMed]
8. Butz, T.; Thiran, J.P. From error probability to information theoretic (multi-modal) signal processing. *Elsevier Signal Process.* **2005**, *85*, 875–902. [CrossRef]
9. Kim, J.; Fisher, J.W., III; Yezzi, A.; Cetin, M.; Willsky, A.S. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. Image Process.* **2005**, *14*, 1486–1502. [PubMed]
10. Padmanabhan, M.; Dharanipragada, S. Maximizing information content in feature extraction. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 512–519. [CrossRef]
11. Silva, J.; Narayanan, S. Minimum probability of error signal representation. Presented at IEEE Workshop Machine Learning for Signal Processing, Thessaloniki, Greece, 27–29 August 2007; pp. 348–353.
12. Silva, J.; Narayanan, S. Discriminative wavelet packet filter bank selection for pattern recognition. *IEEE Trans. Signal Process.* **2009**, *57*, 1796–1810. [CrossRef]
13. Gokcay, E.; Principe, J.C. Information theoretic clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 158–171. [CrossRef]
14. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Stehlik, M. Generalized skew-normal negentropy and its applications to fish condition time factor time series. *Entropy* **2017**, *19*, 528. [CrossRef]
15. Lake, D.E. Nonparametric entropy estimation using kernel densities. *Methods Enzymol.* **2009**, *467*, 531–546. [PubMed]
16. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000; Volume 3.
17. Wu, Y.; Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* **2016**, *62*, 3702–3720. [CrossRef]
18. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885. [CrossRef] [PubMed]
19. Paninski, L. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inf. Theory* **2004**, *50*, 2200–2203. [CrossRef]
20. Valiant, G.; Valiant, P. Estimating the unseen: An $n/log(n)$-sample estimator for entropy and support size, shown opitmal via new CLTs. In Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, San Jose, AL, USA, 6–8 June 2011; pp. 685–694.
21. Valiant, G.; Valiant, P. *A CLT and Tight Lower Bounds for Estimating Entropy*; Technical Report TR 10-179; Electronic Colloquium on Computational Complexity: Potsdam, Germany, 2011; Volume 17, p. 9.
22. Braess, D.; Forster, J.; Sauer, T.; Simon, H.U. How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution. In Proceedings of the International Conference on Algorithmic Learning Theory, Lübeck, Germany, 24–26 November 2004; Springer: Berlin/Heidelberg, Germany, 2002; pp. 380–394.
23. Csiszár, I.; Shields, P.C. Information theory and statistics: A tutorial. In *Foundations and Trends® in Communications and Information Theory*; Now Publishers Inc.: Breda, The Netherlands, 2004; pp. 417–528.
24. Ho, S.W.; Yeung, R.W. On the discontinuity of the Shannon information measures. *IEEE Trans. Inf. Theory* **2009**, *55*, 5362–5374.
25. Silva, J.; Parada, P. Shannon entropy convergence results in the countable infinite case. In Proceedings of the International Symposium on Information Theory, Cambridge, MA, USA, 1–6 July 2012; pp. 155–159.

26. Ho, S.W.; Yeung, R.W. The interplay between entropy and variational distance. *IEEE Trans. Inf. Theory* **2010**, *56*, 5906–5929. [CrossRef]
27. Harremoës, P. Information topologies with applications. In *Entropy, Search, Complexity*; Csiszár, I., Katona, G.O.H., Tardos, G., Eds.; Springer: New York, NY, USA, 2007; Volume 16, pp. 113–150.
28. Devroye, L.; Lugosi, G. *Combinatorial Methods in Density Estimation*; Springer: New York, NY, USA, 2001.
29. Barron, A.; Györfi, L.; van der Meulen, E.C. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inf. Theory* **1992**, *38*, 1437–1454. [CrossRef]
30. Antos, A.; Kontoyiannis, I. Convergence properties of fucntionals estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [CrossRef]
31. Piera, F.; Parada, P. On convergence properties of Shannon entropy. *Probl. Inf. Transm.* **2009**, *45*, 75–94. [CrossRef]
32. Berlinet, A.; Vajda, I.; van der Meulen, E.C. About the asymptotic accuracy of Barron density estimates. *IEEE Trans. Inf. Theory* **1998**, *44*, 999–1009. [CrossRef]
33. Vajda, I.; van der Meulen, E.C. Optimization of Barron density estimates. *IEEE Trans. Inf. Theory* **2001**, *47*, 1867–1883. [CrossRef]
34. Lugosi, G.; Nobel, A.B. Consistency of data-driven istogram methods for density estimation and classification. *Ann. Stat.* **1996**, *24*, 687–706.
35. Silva, J.; Narayanan, S. Information divergence estimation based on data-dependent partitions. *J. Stat. Plan. Inference* **2010**, *140*, 3180–3198. [CrossRef]
36. Silva, J.; Narayanan, S.N. Nonproduct data-dependent partitions for mutual information estimation: Strong consistency and applications. *IEEE Trans. Signal Process.* **2010**, *58*, 3497–3511. [CrossRef]
37. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
38. Gray, R.M. *Entropy and Information Theory*; Springer: New York, NY, USA, 1990.
39. Kullback, S. A lower bound for discrimination information in terms of variation. *IEEE Trans. Inf. Theory* **1967**, *13*, 126–127. [CrossRef]
40. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
41. Kemperman, J. On the optimum rate of transmitting information. *Ann. Math. Stat.* **1969**, *40*, 2156–2177. [CrossRef]
42. Breiman, L. *Probability*; Addison-Wesley: Boston, MA, USA,1968.
43. Varadhan, S. *Probability Theory*; American Mathematical Society: Providence, RI, USA, 2001.
44. Györfi, L.; Páli, I.; van der Meulen, E.C. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inf. Theory* **1994**, *40*, 267–271. [CrossRef]
45. Rissanen, J. *Information and Complexity in Statistical Modeling*; Springer: New York, NY, USA, 2007.
46. Boucheron, S.; Garivier, A.; Gassiat, E. Coding on countably infinite alphabets. *IEEE Trans. Inf. Theory* **2009**, *55*, 358–373. [CrossRef]
47. Silva, J.F.; Piantanida, P. The redundancy gains of almost lossless universal source coding over envelope families. In Proceedings of the IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 2003–2007.
48. Silva, J.F.; Piantanida, P. Almost Lossless Variable-Length Source Coding on Countably Infinite Alphabets. In Proceedings of the IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 1–5.
49. Nobel, A.B. Histogram regression estimation using data-dependent partitions. *Ann. Stat.* **1996**, *24*, 1084–1105. [CrossRef]
50. Silva, J.; Narayanan, S. Complexity-regularized tree-structured partition for mutual information estimation. *IEEE Trans. Inf. Theory* **2012**, *58*, 940–1952. [CrossRef]
51. Darbellay, G.A.; Vajda, I. Estimation of the information by an adaptive partition of the observation space. *IEEE Trans. Inf. Theory* **1999**, *45*, 1315–1321. [CrossRef]

52. Devroye, L.; Györfi, L.; Lugosi, G. *A Probabilistic Theory of Pattern Recognition*; Springer: New York, NY, USA, 1996.
53. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: New York, NY, USA, 2009.