Convergence Properties of Functional Estimates for Discrete Distributions

András Antos * Ioannis Kontoyiannis †

September 26, 2001

ABSTRACT: Suppose P is an arbitrary discrete distribution on a countable alphabet $\mathcal{X}$. Given an i.i.d. sample $(X_1, \ldots, X_n)$ drawn from P, we consider the problem of estimating the entropy $H(P)$ or some other functional $F = F(P)$ of the unknown distribution P. We show that, for additive functionals satisfying mild conditions (including the cases of the mean, the entropy, and mutual information), the plug-in estimates of F are universally consistent. We also prove that, without further assumptions, no rate-of-convergence results can be obtained for *any* sequence of estimators. In the case of entropy estimation, under a variety of different assumptions, we get rate-of-convergence results for the plug-in estimate and for a nonparametric estimator based on match-lengths. The behavior of the variance and the expected error of the plug-in estimate is shown to be in sharp contrast to the finite-alphabet case. A number of other important examples of functionals are also treated in some detail.

Key Words: Functional estimation; entropy estimation; rates of convergence, match lengths.

*A. Antos is with the Informatics Laboratory, Research Division, Computer and Automation Research Institute of the Hungarian Academy of Sciences, H-1518 Lágymányosi u.11, Budapest, Hungary. (email: `antos@szit.bme.hu`).

†I. Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Box F, 182 George St., Providence, RI 02912, USA. (email: `yiannis@dam.brown.edu`). Research supported in part by NSF grants #0073378-CCR and DMS-9615444.

1 Introduction

Suppose X is a discrete random variable with an unknown distribution $P = \{p(i) \; ; \; i \in \mathcal{X}\}$ on the countable alphabet $\mathcal{X}$, and let $F = F(P)$ be an extended-real-valued functional on the space of probability distributions on $\mathcal{X}$ (we allow $F(P)$ to be infinite for some P).

Given n independent and identically distributed (i.i.d.) observations $(X_1, \ldots, X_n)$ drawn from the same distribution as X, it is often important to be able to estimate F accurately, that is, to have an estimate $F_n = F_n(X_1, \ldots, X_n)$ of F such that the difference $|F_n - F|$ is small.

The first question we ask is whether universal estimates exist, that is, whether it is possible to come up with a sequence $\{F_n\}$ of *estimators* for F, such that the difference $|F_n - F|$ converges to zero for all possible distributions P. The second and main question we consider is whether (and under what conditions) it is possible to obtain universal convergence rates for a specific class of estimators.

We show that the natural plug-in estimates are universally consistent for a wide class of additive functionals [recall that the plug-in estimate for a functional $F = F(P)$ is given by $F_n = F(p_n)$, where p_n is the empirical distribution induced on $\mathcal{X}$ by the samples $(X_1, \ldots, X_n)$]. On the other hand, we prove that for a general class of functionals F there is no method that guarantees a certain rate of convergence for all distributions P with $F(P) < \infty$, that is, the convergence of the error of *any* sequence of estimators can be arbitrary slow.

Two important special cases that partly motivated this study are when F is the entropy, and when F is the mutual information. The case of the entropy is particularly interesting, especially in view of the recent attention to the problem of universal compression of memoryless sources with large or infinite alphabets; see [27], [25, p.2061] and the references therein.

Entropy. For a random variable X with distribution $P = \{p(i) \; ; \; i \in \mathcal{X}\}$, the entropy of X is defined by:

$$F = H \stackrel{\triangle}{=} -\sum_{i \in \mathcal{X}} p(i) \log_2 p(i) = \mathbf{E}\{-\log_2 p(X)\}.$$

(Throughout the paper we write log for the natural logarithm and $\log_2$ for the logarithm taken to base 2.)

Mutual information. For two random variables (V, W) with joint distribution $\{p(i,j) \; ; \; i, j \in \mathcal{X}\}$ and marginal distributions $\{p_V(i)\}$ and $\{p_W(j)\}$, the mutual information between V and W is defined by:

$$F = I \stackrel{\triangle}{=} \sum_{(i,j) \in \mathcal{X}^2} p(i,j) \log_2 \frac{p(i,j)}{p_V(i) p_W(j)}.$$

See [9] for an extensive information-theoretic introduction to these and many other entropy-related quantities.

The rest of the paper is organized as follows. In the next section we give a set of mild conditions under which the plug-in estimates are universally consistent, and present specific examples of functionals (including the entropy and mutual information) satisfying these conditions. Section 3 shows that, for a wide class of functionals, there is no universal convergence rate for any sequence of estimators. This is illustrated through a number of examples: There are no universal convergence-rates for entropy estimation, for estimating mutual information, or for Rényi entropies of order $a \in (0,1)$. But this is not always the case: We also show that there do exist functionals F for which it is possible to obtain universal convergence rates (see the example of power-sums in Section 2).

In Sections 4 and 5 we focus on the problem of entropy estimation, a problem which turns out to lead to some somewhat unexpected results. Since no universal convergence rates exist, we consider restricted classes of distributions within which rates of convergence can actually be obtained. First we recall that, in the case when $\mathcal{X}$ is a finite alphabet, the plug-in entropy-estimates $\widehat{H}_n$ converge to the true entropy H at a rate of $\approx \sigma/\sqrt{n}$, where σ^2 is the variance $\sigma^2 = \mathbf{Var}\{-\log_2 p(X)\}$. This suggests that, as long as

$$\sigma^2 = \mathbf{Var}\{-\log_2 p(X)\} < \infty, \tag{1}$$

the same convergence rate might also hold in the infinite-alphabet case. As we show in Corollary 5, this is not at all the case. Under restrictions on the tails of the distributions considered, detailed results about the convergence of the plug-in estimates are given in Theorem 7. In fact we show that there are "many" distributions with $\sigma^2 < \infty$, for which the plug-in estimates converge no faster than $(\log n)^{-2-\epsilon}$; see Corollary 5. [Similar results are proved for the case of mutual information.]

In Section 5 we consider a sequence of entropy estimators based on match-lengths. These estimators are motivated by the Lempel-Ziv family of data compression algorithms, and they are nonparametric in flavor (in that they do not form an estimate of the source distribution and are consistent for arbitrary stationary and ergodic processes). For these estimators we provide positive convergence-rate results under assumptions weaker than those used for the plug-in estimates in the previous section. Specifically, for distributions satisfying the variance condition (1) above (or some variant of this condition) we show that the match-length estimators converge at a rate of order $(\log n)^{-1/2}$. Although rather slow, this rate is obtained under the milder condition (1), under which we were unable to provide upper bounds for the convergence of the plug-in estimates.

Finally, in the Appendix we collect the statements and proofs of several technical lemmas.

2 Consistency of the Plug-In Estimates

Here we prove a general consistency result for plug-in estimates. Let X be a discrete random variable with distribution $\{p(i)\ ;\ i \in \mathcal{X}\}$ on $\mathcal{X}$. We consider a class of functionals that we call *additive functionals*, given by the general form

$$F \stackrel{\triangle}{=} g\left(\sum_{i\in\mathcal{X}} f(i, p(i))\right).$$

Here, f and g are arbitrary real-valued functions, with the only restriction that f is always non-negative. The *plug-in estimate* for F is defined by

$$\widehat{F}_n \stackrel{\triangle}{=} g\left(\sum_{i\in\mathcal{X}} f(i, p_n(i))\right),$$

where

$$p_n(i) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j=i\}}$$

is the empirical distribution induced by the samples $(X_1, \ldots, X_n)$ on $\mathcal{X}$. In other words, $\widehat{F}_n = F(p_n)$.

For each $i \in \mathcal{X}$ let f_i denote the function $f(i, \cdot) : [0,1] \to \mathcal{R}$.

Theorem 1 *Assume that for some positive constants p and K, for each $i \in \mathcal{X}$, and all integers $0 \le j < n$, f_i satisfies*

$$\left| f_i\left(\frac{j+1}{n}\right) - f_i\left(\frac{j}{n}\right)\right| \le \frac{K}{n^p}.$$

If, in addition, g is Lipschitz(q, K_0) (where q, $K_0 > 0$), that is,

$$|g(x) - g(y)| \le K_0|x-y|^q, \quad x, y \in \mathcal{R},$$

then for every $\epsilon > 0$,

$$\mathbf{P}\left\{|\widehat{F}_n - \mathbf{E}(F_n)| > \epsilon\right\} \le 2e^{-K_1 n^{2pq-1}\epsilon^2} \tag{2}$$

and

$$\mathbf{Var}(\widehat{F}_n) \le \frac{1}{2K_1 n^{2pq-1}}, \tag{3}$$

where $K_1 = 2/[K_0^2(2K)^{2q}]$. Moreover, if $pq > 1/2$, then

$$\lim_{n\to\infty}[\widehat{F}_n - \mathbf{E}(\widehat{F}_n)] = 0 \quad \text{almost surely (a.s.)} \quad \text{and} \quad \lim_{n\to\infty}\mathbf{Var}(\widehat{F}_n) = 0.$$

For the proof of Theorem 1 we will need the following two propositions, as well as a number of simple lemmas, stated and proved in the Appendix. The first proposition is a version of Azuma's inequality; see, e.g., [19], [16], or [24].

Proposition 1 *Let $X_1, \ldots, X_n$ be independent random variables on $\mathcal{X}$, and assume that $\widehat{F} : \mathcal{X}^n \to \mathcal{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x'_j\in\mathcal{X}} |\widehat{F}(x_1, \ldots, x_n) - \widehat{F}(x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_n)| \le c_j \quad , \quad 1 \le j \le n.$$

Then for any $\epsilon > 0$,

$$\mathbf{P}\{\widehat{F}(X_1, \ldots, X_n) - \mathbf{E}[\widehat{F}(X_1, \ldots, X_n)] \ge \epsilon\} \le e^{-2\epsilon^2/\sum_{j=1}^n c_j^2}.$$

The following proposition is given by Devroye in [13].

Proposition 2 *If the conditions of Proposition 1 hold, then*

$$\mathbf{Var}\{\widehat{F}(X_1,\ldots,X_n)\} \le \frac{1}{4}\sum_{j=1}^{n} c_j^2.$$

PROOF OF THEOREM 1. Proposition 1 implies (2). To see this, note that by changing the value of one sample point X_j, there can be two values i' and i'' such that $p_n(i')$ increases by $1/n$ and $p_n(i'')$ decreases by the same amount. Then by the properties of f_i, the value of $\sum_i f_i(p_n(i))$ cannot change by more than $2K/n^p$, hence, by the properties of the function g, the value of $\widehat{F}_n$ cannot change by more than

$$\frac{K_0(2K)^q}{n^{pq}}.$$

Similarly Proposition 2 implies (3).

For $pq > 1/2$, equation (2) and the Borel-Cantelli lemma imply $\lim_{n\to\infty}(\widehat{F}_n - \mathbf{E}[\widehat{F}_n]) = 0$ almost surely (a.s.), and (3) implies the other statement. □

Theorem 2 *Suppose f and g satisfy the conditions of Theorem 1, with $pq > 1/2$. If g is concave and monotone increasing, and all the functions f_i are continuous and they satisfy*

$$\limsup_{n\to\infty}\sum_i \mathbf{E}[f_i(p_n(i))] \le \sum_i f_i(p(i)), \tag{4}$$

then $\{\widehat{F}_n\}$ is strongly universally consistent, that is,

$$\lim_{n\to\infty}\widehat{F}_n = F \quad a.s.$$

If $F < \infty$, it is also consistent in L^2, that is,

$$\lim_{n\to\infty}\mathbf{E}\{(\widehat{F}_n - F)^2\} = 0.$$

Note that if all the f_i are concave, then (4) is satisfied (Jensen's inequality).

PROOF. By Lemma 6 of the Appendix and Theorem 1,

$$\lim_{n\to\infty}\widehat{F}_n = \lim_{n\to\infty}(\widehat{F}_n - \mathbf{E}[\widehat{F}_n]) + \lim_{n\to\infty}\mathbf{E}[\widehat{F}_n] = F \quad \text{a.s.}$$

To get the L^2 consistency for $F < \infty$ observe that

$$\mathbf{E}\{(\widehat{F}_n - F)^2\} = \mathbf{Var}(\widehat{F}_n) + (F - \mathbf{E}[\widehat{F}_n])^2 \to 0 \quad (n\to\infty). \ \square \tag{5}$$

Examples:

If the support of X is finite, then Lemma 7 of the Appendix implies strong and L^2 consistency for all of the functionals below. More generally, when $\mathcal{X}$ is countably infinite:

Expectation. (Here $\mathcal{X} \subset \mathcal{R}$.) Taking $f(i, p(i)) = ip(i)$ and $g(x) = x$, Lemma 7 of the Appendix (or Theorem 2) implies the strong law of large numbers for bounded, discrete random variables. [But note that we already used the strong law for the empirical distribution in the proofs!] Similarly, the moments of X are additive functionals, and thus, for example, the variance is the sum of two additive functionals.

Entropy. The plug-in estimate of the entropy H is

$$\widehat{H}_n = -\sum_{i \in \mathcal{X}} p_n(i) \log_2 p_n(i).$$

Corollary 1 *The plug-in estimate of H is strongly universally consistent, that is,*

$$\lim_{n\to\infty} \widehat{H}_n = H \quad a.s.$$

For $H < \infty$ it is also consistent in L^2, that is,

$$\lim_{n\to\infty} \mathbf{E}\{(\widehat{H}_n - H)^2\} = 0.$$

Proof. Clearly H is an additive functional with $f(i, p(i)) = -p(i)\log_2 p(i)$ and $g(x) = x$. The conditions of Theorems 1 and 2 are satisfied with $q = 1$, $K_0 = 1$, with any $1/2 < p < 1$ and $K = \frac{1}{e(1-p)\log 2}$. □

Remark. Using methods similar to the ones used to prove Theorem 1, the following additional properties of $\widehat{H}_n$ can also be shown to hold:

(i) Since $\widehat{H}_n$ is the entropy of a distribution concentrating on at most n different points (namely, p_n), it is $0 \le \widehat{H}_n \le \log_2 n$, for all n.

(ii) For all n, $\mathbf{E}[\widehat{H}_n] \le H$.

(iii) For every n and $\epsilon > 0$,

$$\mathbf{P}\left\{|\widehat{H}_n - \mathbf{E}(\widehat{H}_n)| > \epsilon\right\} \le 2e^{-n\epsilon^2/2\log_2^2 n},$$

because for all integers $0 \le j < n$,

$$\left|\frac{j+1}{n}\log_2\frac{j+1}{n} - \frac{j}{n}\log_2\frac{j}{n}\right| \le \frac{\log_2 n}{n}.$$

(iv) For all n, $\mathbf{Var}(\widehat{H}_n) \le (\log_2^2 n)/n$.

Mutual information. Let

$$p_n(i,j) = \frac{1}{n}\sum_{k=1}^{n} I_{\{V_k=i, W_k=j\}}$$

denote the empirical distribution induced on $\mathcal{X}^2$ by the i.i.d. samples (V_k, W_k), $k = 1\ldots,n$, and let $\{p_{V,n}(i)\}$ and $\{p_{W,n}(j)\}$ be the corresponding marginals. Using the identity (see, e.g., [9])

$$I = H(V) + H(W) - H(V,W),$$

the results for entropy estimation imply the universal consistency of the plug-in estimate

$$\widehat{I}_n = \sum_{(i,j)\in\mathcal{X}^2} p_n(i,j)\log_2 \frac{p_n(i,j)}{p_{V,n}(i)p_{W,n}(j)}.$$

Corollary 2 *If $H(V,W)$ is finite, then the plug-in estimate of I is strongly universally consistent and consistent in L^2, that is,*

$$\lim_{n\to\infty}\widehat{I}_n = I \quad a.s. \quad and \quad \lim_{n\to\infty}\mathbf{E}\{(\widehat{I}_n - I)^2\} = 0.$$

Note, however, that it is possible to have $H(V,W) = \infty$ while $I < \infty$. Also note that the almost sure consistency of $\widehat{H}_n$ and $\widehat{I}_n$ is valid even for general stationary and ergodic sources.

Sum of i-dependent powers. Let $\mathcal{X} = \{2,3,4,\ldots\}$ and define

$$F = \sum_{i=2}^{\infty} p^{i/(i+1)}(i).$$

Clearly F is an additive functional, with $f(i,p(i)) = p^{i/(i+1)}(i)$ and $g(x) = x$. It is easy to check that the conditions of Theorems 1 and 2 are satisfied with $q = 1$, $K_0 = 1$, $p = 2/3$ and $K = 1$. Therefore, $\widehat{F}_n$ is strongly universally consistent and consistent in L^2.

Power sums. For any $a > 0$,

$$F^{(a)} \triangleq \sum_{i\in\mathcal{X}} p^a(i)$$

is an additive functional with $f(i,p(i)) = p^a(i)$ and $g(x) = x$. Note that (with the convention that $0^0 = 0$), $F^{(0)}$ is simply the size of the support of the distribution $\{p(i)\}$. In this case it is trivial to check the strong and L^p universal consistency of $\{\widehat{F}_n^{(0)}\}$. For $a > 0$ we consider two cases separately.

Case 1: $0 < a \le 1$. Note that $\widehat{F}_n^{(a)} \in [1, n^{1-a}]$. The conditions of Theorem 1 are satisfied with $q = 1$, $K_0 = 1$, $p = a$ and $K = 1$, and also those of Theorem 2, if $pq = a > 1/2$. Therefore, for $a > 1/2$, $\widehat{F}_n^{(a)}$ is universally consistent almost surely and in L^2. For $a \le 1/2$ we get that $\{\widehat{F}_n^{(a)}\}$ is asymptotically unbiased, but for the variance we only get that it is of order $O(n^{1-2a})$. [For $a = 0$, it is easy to check that $\mathbf{Var}(\widehat{F}_n^{(0)}) \to 0$.]

If we change our strategy and bound the fluctuations of $\widehat{F}_n^{(a)}$ on the average (instead of the worst-case) when changing only one sample, then it is possible to use an Efron-Stein type inequality of

Steele [23] in place of Proposition 2, leading to the bound $\mathbf{Var}(\widehat{F}_n^{(a)}) = o(n^{1-2a})$. This implies L^2 consistency for $a = 1/2$. But it is possible to get much more: The strong and L^2 consistency of $\widehat{F}_n^{(a)}$ for any $0 < a \leq 1$ can be obtained by replacing Azuma's inequality by the following concentration inequality due to Boucheron, Lugosi and Massart [7]:

Proposition 3 *Let $(X_1, \ldots, X_n)$ be independent random variables on $\mathcal{X}$. For some $\widehat{F} : \mathcal{X}^n \to [0, \infty)$ assume that there exists a function $G : \mathcal{X}^{n-1} \to \mathcal{R}$ such that, for any $x_1, \ldots, x_n \in \mathcal{X}$:*

(1) $0 \leq \widehat{F}(x_1, \ldots, x_n) - G(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \leq 1\,, \quad 1 \leq j \leq n;$

(2) $\sum_{j=1}^n (\widehat{F}(x_1, \ldots, x_n) - G(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)) \leq \widehat{F}(x_1, \ldots, x_n).$

Then for any $\epsilon > 0$,

$$\mathbf{P}\{|\widehat{F}(X_1, \ldots, X_n) - \mathbf{E}[\widehat{F}(X_1, \ldots, X_n)]| \geq \epsilon\} \leq 2e^{-\epsilon^2/(2\mathbf{E}(\widehat{F})+2\epsilon/3)}.$$

The proof of the following Corollary is in the Appendix.

Corollary 3 *For $0 \leq a \leq 1$ and any $\epsilon > 0$,*

$$\mathbf{P}\{|\widehat{F}_n^{(a)} - \mathbf{E}[\widehat{F}_n^{(a)}]| \geq \epsilon\} \leq 2e^{-n^a \epsilon^2/(2\mathbf{E}[\widehat{F}_n^{(a)}]+2\epsilon/3)}. \tag{6}$$

Since $\mathbf{E}[\widehat{F}_n^{(a)}] \to F^{(a)}$, this inequality and the Borel-Cantelli lemma imply $\lim_{n\to\infty} \widehat{F}_n^{(a)} = F^{(a)}$ a.s. It is now a straightforward calculation to deduce from (6) that

$$\mathbf{Var}(\widehat{F}_n^{(a)}) = O(n^{-a}),$$

also proving the L^2 consistency for any $0 < a \leq 1$.

Case 2: $a > 1$. Note that $\widehat{F}_n^{(a)} \in [n^{1-a}, 1]$, and that f_i is convex for each i. Assume for simplicity that a is integer. The following is proved in the Appendix:

Lemma 1 *For positive integer a,*

$$\sum_i \mathbf{E}\{p_n^a(i)\} \leq O(1/n) + \sum_i p^a(i), \tag{7}$$

and thus $\widehat{F}_n^{(a)}$ satisfies (4).

Hence the conditions of Theorems 1 and 2 are satisfied with $q = 1$, $K_0 = 1$, $p = 1$ and $K = a$, proving the strong universal consistency and L^2 consistency of $\widehat{F}_n^{(a)}$. Moreover, we see from (7) and the convexity of f_i that

$$(\mathbf{E}[\widehat{F}_n^{(a)}] - F^{(a)})^2 = O(1/n^2),$$

and by (3), $\mathbf{Var}(\widehat{F}_n^{(a)}) = O(1/n)$, hence using (5) we get that

$$\mathbf{E}\{(\widehat{F}_n^{(a)} - F^{(a)})^2\} = O(1/n). \tag{8}$$

Rényi entropies. The Rényi entropy of order $0 < a < 1$ is

$$F^{(a)} = \frac{\log_2 \left(\sum_{i \in \mathcal{X}} p^a(i)\right)}{1-a},$$

which is an additive functional with $f(i, p(i)) = p^a(i)$ and $g(x) = \log_2 x/(1-a)$ on $[1, \infty)$. Notice that $\widehat{F}_n^{(a)} \in [0, \log_2 n]$ and that $\sum_i p^a(i) \geq 1$. Also we have $F^{(a)} \to H$ and $\widehat{F}_n^{(a)} \to \widehat{H}_n$ as $a \to 1$.

The conditions of Theorem 1 are satisfied with $q = 1$, $K_0 = 1/(1-a)\log 2$, $p = a$ and $K = 1$, and also those of Theorem 2, if $pq = a > 1/2$. Therefore, for $a > 1/2$, we get the strong universal consistency and L^2 consistency of $\widehat{F}_n^{(a)}$. Moreover, using the same arguments as above (based on the Lipschitz property of g), it is not hard to obtain corresponding results for all $0 < a < 1$.

Error functionals in classification. In classification, the plug-in estimate of the Bayes error probability and the plug-in estimate of the asymptotical error probability of the k-nearest neighbor rule, can both be shown to be universally consistent, along the same lines as the proof of Theorem 2; cf. [1] and [16].

In the case of power sums $F^{(a)}$ with $a > 1$ we saw that the plug-in estimate was not only universally consistent, but it also had a universal rate of convergence in L^2 (see equation (8)). It is therefore natural to ask whether the same is true for other additive functionals. In the following section we show that, in general, it is not.

3 Slow Rate of Convergence

Here we show that for a large class of additive functionals there is no universal convergence rate, not only for the plug-in estimates but for *any* sequence of estimators (see Corollary 4 below). Similar global slow-rate-of-convergence results have been obtained for pattern recognition [8][10][16], for regression function estimation and density estimation [11][6][14], and for several other functionals [2][3]. In general, we anticipate that universal convergence rates do not exist when the class of distributions considered is sufficiently rich. For example, the absence of universal rates may stem from dependence of the data, or from the absence of restrictions on the tails of the distribution generating the data.

Most of the results here (as well as the results in the literature mentioned above) are based on the use of a "well-separating" subclass of distributions, that is, a collection of distributions which are "close" to one another, but over which the values of the functional of interest are significantly different. (See also [12][15].) In terms of proof technique, usually such results are obtained by using as our well-separating class a "rectangular" class of distributions, parametrized by infinite binary sequences $u = (u_0, u_1, u_2, \ldots)$; the term "rectangular" refers to the fact that the collection of all such sequences u can be thought of as an infinite-dimensional rectangle or "cube." Then the parameter u is chosen randomly, and the worst-case error in this subclass is bounded below by the average error according to this random parameter. More precisely, the proof is based on the fact that for two distributions that lie on a common "edge" of this rectangle, the values of the functional of interest are significantly different, while the data typically (i.e., with high probability) contains little or no information regarding which of the two distributions is the true one. This idea is reflected in the assumptions (and the proof) of Theorem 3 below.

Let F be a given functional (not necessarily additive), let $\mathcal{D}$ be a class of distributions on $\mathcal{X}$, and write $\mathcal{N}$ for the set of nonnegative integers $\{0, 1, 2, \ldots\}$. As before, X denotes a random variable with distribution $P = \{p(i)\ ;\ i \in \mathcal{X}\}$, and we write $D_n = (X_1, \ldots, X_n)$ for a vector of n i.i.d. random variables with distribution P. The next theorem establishes a general lower bound for the rate of convergence of an arbitrary sequence of estimators $\{F_n\}$.

Theorem 3 *Let $d_0 = 0$ and $\{d_i\ ;\ i \geq 1\}$ be a sequence of positive real numbers. Assume that for any discrete weight vector $\{q_0, q_1, \ldots\}$ with $\sum_{i=0}^{\infty} q_i = 1$ and $0 < q_i \leq 2^{-i}$ $(i \geq 1)$, there is a subclass of distributions $\{\mu_u : u \in \{0,1\}^{\mathcal{N}}\} \subseteq \mathcal{D}$ parametrized by binary sequences $u = (u_0, u_1, u_2, \ldots)$, having the following properties:*

1. There are disjoint subsets $B_0, B_1, B_2, \ldots$ of $\mathcal{X}$ such that $\mu_u(B_i) = q_i$ for all u and all i.

2. The restriction of μ_u to B_i is chosen from two possibilities according to the value of u_i.

3. Let $F(u) = F(\mu_u)$. If u and u' are two binary sequences coinciding in all but the k^{th} position, then

$$|F(u) - F(u')| \geq d_k.$$

Assume moreover that $F(u)$ is finite for all u.

Then, for any sequence of estimators $\{F_n\}$ and any sequence $\{a_n\}$ of positive numbers converging to zero, there is a distribution P in $\mathcal{D}$ with $|F| < \infty$, for which

$$\mathbf{P}\{|F_n - F| > a_n\} > \frac{1}{2} - \epsilon \quad \textit{infinitely often, for any } \epsilon > 0.$$

Now applying Theorem 3 to the sequence $\{\sqrt{a_n}\}$ instead of $\{a_n\}$ we get that for any $\{F_n\}$ there is a distribution in $\mathcal{D}$ with $|F| < \infty$ such that for any $K > 0$

$$\limsup_{n\to\infty} \mathbf{P}\{|F_n - F| > K a_n\} \geq \frac{1}{2}.$$

This observation immediately gives us the following general slow-rate result for the expected estimation error:

Corollary 4 *Under the conditions of Theorem 3, for any sequence $\{F_n\}$ of estimators and for any sequence $\{a_n\}$ of positive numbers converging to zero, there is a distribution P in $\mathcal{D}$ with $|F| < \infty$, for which*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|F_n - F|\}}{a_n} = \infty.$$

Remark. The phrase "infinitely often" cannot be dropped from Theorem 3, and similarly the lim sup in Corollary 4 cannot be replaced by lim inf. Indeed, there exist deterministic sequences

$\{f_n\}$ with $|f_n - F| \leq 2/\sqrt{n}$ infinitely often, simultaneously for every F. Just consider the dyadic sequence

$$\{f'_n\} = \left\{\frac{1}{2^0}, \frac{1}{2^1}, \frac{2}{2^1}, \frac{3}{2^1}, \frac{4}{2^1}, \frac{1}{2^2}, \frac{2}{2^2}, \ldots, \frac{16}{2^2}, \frac{1}{2^3}, \frac{2}{2^3}, \ldots, \frac{64}{2^3}, \ldots\right\}$$

and let $f_{2n-1} = f'_n$ and $f_{2n} = -f'_n$. Now for every F, for every i large enough, there is an element of the sequence in the first $2(1+4+\cdots+4^i) < 4\cdot 4^i$ elements, whose distance from F is at most 2^{-i}. We thus obtain a very good estimator along an (unknown) subsequence for every value of F. (This construction can also be generalized to certain finite dimensional spaces; see [6] for details.)

When the samples $D_n = (X_1, \ldots, X_n)$ are generated from the distribution μ_u, we write $D_n(u)$ for D_n to indicate how the samples depend on u, and write $\mu_u^{(n)}$ for the (product) distribution of D_n. The following lemma is the main ingredient in the proof of Theorem 3; its proof is in the Appendix.

Lemma 2 *Consider a class $\{\mu_u : u \in \{0,1\}^{\mathcal{N}}\}$ of distributions parametrized by binary sequences $u = (u_0, u_1, u_2, \ldots)$, and assume that there exist subsets $A_{n,k} \subset \mathcal{X}^n$ such that, if u and u' are two binary sequences coinciding in all but the k^{th} position, then for every string $x_1^n \triangleq (x_1, \ldots, x_n) \in A_{n,k}$*

$$\mu_u^{(n)}(x_1^n) = \mu_{u'}^{(n)}(x_1^n) \tag{9}$$

and

$$|F(u) - F(u')| \geq d_k. \tag{10}$$

Then, for any sequence $\{F_n\}$ of estimators, for any sequence $\{a_n\}$ of positive numbers converging to zero, and for any subsequence $\{n_k\}_{k\in\mathcal{N}}$ of $\{1, 2, \ldots\}$,

$$\sup_u \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u)) - F(u)| > a_n\} \geq \frac{1}{2}\limsup_{k\to\infty}(I_{\{d_k > 2a_{n_k}\}} \inf_u \mu_u^{(n_k)}(A_{n_k,k})).$$

Proof of Theorem 3. To get some intuition, first observe that if none of the samples $X_j \in B_k$, then the estimator has no information about u_k, which gives a contribution to the error of at least d_k. For a sample size n large enough, this is greater than a_n. On the other hand, the probability of this event can be very large, if the measure of B_k is chosen to be small enough. This argument can be made precise as follows:

Given a vector $\{q_i\}$ we will apply Lemma 2 to the sets $A_{n,k} = \{x_1^n : x_j \notin B_k, \text{ for all } 1 \leq j \leq n\}$. Now we have

$$\mu_u^{(n)}(A_{n,k}) = (1 - q_k)^n$$

independently of u, moreover if u and u' differ only in the k^{th} bit and $x_1^n \in A_{n,k}$, then

$$\mu_u^{(n)}(x_1^n) = \mu_{u'}^{(n)}(x_1^n).$$

Hence, by Lemma 2, for any sequence $\{F_n\}$ of estimators, for any sequence $\{a_n\}$ of positive numbers converging to zero, and for any subsequence $\{n_k\}$,

$$\sup_u \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u)) - F(u)| > a_n\} \geq \frac{1}{2}\limsup_{k\to\infty}(I_{\{d_k > 2a_{n_k}\}}(1 - q_k)^{n_k}).$$

Choosing $\{n_k\}$ such that $d_k > 2a_{n_k}$ $(k \geq 1)$, and choosing the distribution vector $\{q_k\}$ such that $q_k = o(1/n_k)$ as $k \to \infty$ (e.g., taking $q_k = \min(a_{n_k}/n_k, 2^{-k})$ and $q_0 = 1 - \sum_{k=1}^{\infty} q_k$), makes the right-hand side above equal to 1/2. So for any sequence $\{a_n\}$ there is $\{q_k\}$ and u such that for any $\epsilon > 0$

$$\mathbf{P}\{|F_n(D_n(u)) - F(u)| > a_n\} > \frac{1}{2} - \epsilon \quad \text{infinitely often.} \quad \square$$

Next we apply Corollary 4 to the cases of the entropy, mutual information, power sums, and Rényi entropies, to obtain the following slow-rate results. For simplicity, in the rest of this section we assume $\mathcal{X} = \mathcal{N} = \{0, 1, 2, \ldots\}$.

Theorem 4 (Entropy) *For any sequence $\{H_n\}$ of estimators for the entropy, and for any sequence $\{a_n\}$ of positive numbers converging to zero, there is a distribution P on $\mathcal{X}$ with $H = H(P) < \infty$ and*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|H_n - H|\}}{a_n} = \infty.$$

Proof. Take $d_k = 2^{-k}$ $(k \geq 1)$. Given $\{q_0, q_1, \ldots\}$ with $0 < q_i \leq 2^{-i}$ $(i \geq 1)$, we pick a sequence $\{l'_0 = 1, l'_1, l'_2, \ldots\}$ of positive integers (to be specified later), and we partition $\mathcal{X}$ into consecutive blocks B_i of cardinalities l'_i. We define $\mu_u(B_i) = q_i$ for all u and i, and given a binary vector u we define μ_u on B_i as follows: If $u_i = 1$ then X is drawn uniformly over the l'_i integers in that block, while if $u_i = 0$, then X takes the value of the first point in the block. Take μ_u to be the distribution of X and let $l_i = \log_2 l'_i$. For this μ_u it is easy to verify that

$$H = H(u) = \sum_{i=1}^{\infty} u_i q_i \log_2 l'_i - \sum_{i=0}^{\infty} q_i \log_2 q_i = \sum_{i=1}^{\infty} u_i q_i l_i - \sum_{i=0}^{\infty} q_i \log_2 q_i.$$

If u and u' differ only in the k^{th} bit, then

$$H(u) - H(u') = \sum_{i=1}^{\infty} (u_i - u'_i) q_i l_i = (u_k - u'_k) q_k l_k,$$

and thus, $|H(u) - H(u')| = q_k l_k$. For $k \geq 1$ the inequality $2^{-k} \leq q_k l_k \leq 2^{-k} + q_k$ is satisfied if, for example, $l'_k = \lceil 2^{1/(q_k 2^k)} \rceil$, and thus $H(u) \leq 4 + \sum_{i=2}^{\infty} i 2^{-i}$. The result follows from Corollary 4. $\square$

Theorem 5 (Mutual Information) *For any sequence $\{I_n\}$ of mutual information estimators, and for any sequence $\{a_n\}$ of positive numbers converging to zero, there are random variables (V, W) with values in $\mathcal{X} \times \mathcal{X}$ such that $I = I(V; W) < \infty$ and*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|I_n - I|\}}{a_n} = \infty.$$

Proof. This is similar to the proof of Theorem 4, using the following construction: Partition $\mathcal{X}$ into consecutive blocks B_i' of cardinalities l_i'. Let $B_i = B_i' \times B_i'$, take $\mu_u(B_i) = q_i$, and for a binary vector u define the distribution μ_u of (V, W) on B_i as follows: If $u_i = 1$ then V is drawn uniformly over the l_i' integers in that block and $W = V$, while if $u_i = 0$ then V and W are drawn uniformly and independently over the l_i' integers in that block. The rest of the proof follows along the same lines as above. □

Theorem 6 (Power sums) *Let $0 < a < 1$. For any sequence $\{F_n^{(a)}\}$ of estimators and any sequence $\{a_n\}$ of positive numbers converging to zero, there is a distribution P on $\mathcal{X}$ such that $F^{(a)} < \infty$ and*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|F_n^{(a)} - F^{(a)}|\}}{a_n} = \infty.$$

Proof. Take $d_k = 2^{-k}$ $(k \geq 1)$. Given $\{q_0, q_1, \ldots\}$ with $0 < q_i \leq 2^{-i}$ $(i \geq 1)$, we pick a sequence $\{l_0' = 1, l_1', l_2', \ldots\}$ of positive integers (to be specified later) and partition $\mathcal{X}$ into consecutive blocks B_i of cardinalities l_i'. We define $\mu_u(B_i) = q_i$ for all u and i, and given a binary vector u we define μ_u on B_i as follows: If $u_i = 1$ then X is drawn uniformly over the l_i' integers in that block, while if $u_i = 0$, then X takes the value of the first point in the block. Take μ_u to be the distribution of X and let $l_i = (l_i')^{1-a} - 1$. For this distribution it is easy to verify that

$$F^{(a)} = F^{(a)}(u) = \sum_{i=1}^{\infty} u_i q_i^a ((l_i')^{1-a} - 1) + \sum_{i=0}^{\infty} q_i^a = \sum_{i=1}^{\infty} u_i q_i^a l_i + \sum_{i=0}^{\infty} q_i^a.$$

If u and u' differ only in the k^{th} bit, then

$$F^{(a)}(u) - F^{(a)}(u') = \sum_{i=1}^{\infty} (u_i - u_i') q_i^a l_i = (u_k - u_k') q_k^a l_k,$$

and thus, $|F^{(a)}(u) - F^{(a)}(u')| = q_k^a l_k$. For $k \geq 1$ the inequality $2^{-k} \leq q_k^a l_k \leq 2^{-k} + q_k^a$ is satisfied if, for example, $l_k' = \lceil (q_k^{-a} 2^{-k} + 1)^{1/(1-a)} \rceil$, and thus $F^{(a)}(u) \leq 2 + 2\sum_{i=1}^{\infty} 2^{-ia}$. The result follows from Corollary 4. □

Remark. In the special case $a = 1$ we have $\widehat{F}_n^{(1)} \equiv F^{(1)} \equiv 1$, hence $\widehat{F}_n^{(1)} - F^{(1)} \equiv 0$. In the case $a = 0$ and $F^{(0)} < \infty$, $F^{(0)} - \widehat{F}_n^{(0)} = \#\{i : p(i) > 0, \nexists j \;\; X_j = i\}$, so

$$\mathbf{E}\{|\widehat{F}_n^{(0)} - F^{(0)}|\} = \sum_{i:p(i)>0} (1 - p(i))^n = O(e^{-\alpha n}),$$

where $\alpha = \min_{i:p(i)>0} \log[1/(1 - p(i))] > 0$.

Rényi entropies. The slow-rate-of-convergence results for the power sums imply analogous slow-rate-of-convergence results for the Rényi entropies in the case $0 < a < 1$. As above, in the special case $a = 0$ and $F^{(0)} < \infty$ we get a universal rate of convergence for $\widehat{F}_n^{(0)}$, of order $O(e^{-\alpha n})$.

REMARK. Theorem 3 can also be applied to obtain corresponding slow-rate-of-convergence results on the estimation of the expectation (cf. [2]). A more general version of Lemma 2 can be applied to obtain analogous slow-rate-of-convergence results for the Bayes error probability and the asymptotic error probability of the k-nearest neighbor rule in classification [1][3][16], and for the optimal error in regression function estimation.

4 Entropy Estimation: The Plug-in Estimates

In Section 2 we saw that, for $a = 2, 3, \ldots$, the functionals $F^{(a)} = \sum_{i\in\mathcal{X}} p^a(i)$ can be consistently estimated, and that the L^2-error of the plug-in estimate is of order $O(1/n)$. But for a wide class of other functionals, including the entropy, we showed that universal convergence rates *cannot* be obtained for *any* sequence of estimators. Therefore, for positive rate-of-convergence results, additional conditions need to be placed on the class of distributions we consider.

In this and the following section we concentrate on the entropy, although some results are given for the case of mutual information. [Note also that there are sharp rate-of-convergence results for the expectation – see [4][20, Theorem 2.6.20][2].]

4.1 Heuristics

Finite Alphabets. In the finite-alphabet case, a relatively straightforward calculation shows that the plug-in estimate $\widehat{H}_n$ is asymptotically Normal,

$$\sqrt{n}[\widehat{H}_n - H] \to N(0, \sigma^2) \quad \text{in distribution,}$$

where $\sigma^2 = \mathbf{Var}\{-\log_2 p(X)\}$; see, e.g., [5]. In particular, for $\sigma^2 > 0$,

$$\mathbf{E}\{|\widehat{H}_n - H|\} = \Theta\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \mathbf{E}\{(\widehat{H}_n - H)^2\} = \Theta\left(\frac{1}{n}\right).$$

This suggests that we should perhaps expect corresponding results in the infinite-alphabet case as long as $\sigma^2 < \infty$, or at least when some of the higher moments $H^{(r)} \triangleq \mathbf{E}\{(-\log_2 p(X))^r\}$ are finite (for some $r > 2$). Somewhat surprisingly, Theorem 7 below shows that this is not at all the case.

Slow Rates. Next we give a heuristic argument based on the proof of our slow-rate result (Theorem 4), indicating what type of conditions we might need to consider for positive rate-of-convergence results. In the proof of Theorem 4, instead of taking $d_k = 2^{-k}$, we could have used any positive sequence of d_k's with $d_0 = 0$ and $\sum_k d_k < \infty$ (moreover, $\{d_k\}$ may depend on the sequence $\{a_n\}$), and taken $l_k = \lceil d_k/q_k \rceil$ (assuring $d_k \le q_k l_k \le d_k + q_k$). This would allow us to choose n_k in the proof of Theorem 3 satisfying only $\sum_k a_{n_k} < \infty$, and then, for example, let $d_k = 3a_{n_k}$, $q_k = o(1/n_k)$.

Now examine the moment parameter

$$H^{(r)} = \mathbf{E}\{(-\log_2 p(X))^r\} = \sum_i p(i)\log_2^r(1/p(i)), \quad (r \ge 1)$$

where $H^{(1)}$ is of course just the entropy. For a distribution corresponding to u in the above construction, and some $r \ge 1$,

$$\sum_k q_k u_k l_k^r \le H^{(r)}(u) = \sum_k q_k(-\log_2 q_k + u_k l_k)^r \le 2^{r-1}\left(\sum_k q_k \log_2^r \frac{1}{q_k} + \sum_k q_k u_k l_k^r\right).$$

Thus, by a rough calculation, for $r > 1$, $H^{(r)}$ is finite here for all u if and only if

$$\sum_k q_k l_k^r = \sum_k (q_k l_k)^r / q_k^{r-1} \asymp \sum_k a_{n_k}^r n_k^{r-1}/o(1) + \sum_k o(1)/n_k < \infty.$$

The above series can be made finite by the choice of $\{n_k\}$ if and only if $\liminf_{n\to\infty} a_n^r n^{r-1} = 0$, that is, if $a_n \ne \Omega(n^{-\frac{r-1}{r}})$. This suggests that maybe the rate $\{n^{-\frac{r-1}{r}}\}$ can be achieved if we restrict ourselves to the case $H^{(r)} < \infty$, at least for $1 < r \le 2$.

In the next subsection we give rate-of-convergence results for the plug-in estimate of the entropy, under assumptions on the tail of the distribution of X. In the following subsection we give upper and lower bounds for the convergence of a nonparametric estimator based on match-lengths, under assumptions on the finiteness of $H^{(r)}$. As we will see, neither the plug-in nor the match-length estimators achieve the $O(n^{-\frac{r-1}{r}})$-rate suggested above.

4.2 Tail conditions

Here we prove a sharp rate-of-convergence result for the plug-in estimates of entropy, assuming appropriate tail conditions. Instead of the finiteness of the r^{th} moment of $[-\log p(X)]$, we restrict our attention to (the smaller class of) distributions with tail probabilities decreasing approximately like (const)i^{-q}, for some $q > 1$. In this subsection, without loss of generality we take $i \in \mathcal{X} = \mathcal{N}$ to always be a nonnegative integer. The following theorem shows that, under appropriate tail conditions, the L^1 error of the plug-in estimate for the entropy is exactly $\Theta(n^{-\frac{q-1}{q}})$ for $q < 2$.

Theorem 7 *Assume that for some $q > 1$ there exist positive constants c_1, $c_2 > 0$ such that $c_1/i^q \le p(i) \le c_2/i^q$, $i = 1, 2, \ldots$. Then, for the plug-in estimate, if $q \in (1, 2)$,*

$$\Omega\left(n^{-\frac{q-1}{q}}\right) = \mathbf{E}\{|\widehat{H}_n - H|\} \le (\mathbf{E}\{(\widehat{H}_n - H)^2\})^{1/2} = O\left(n^{-\frac{q-1}{q}}\right).$$

The same result holds for $q \ge 2$, with the last upper bound above replaced by

$$(\mathbf{E}\{(\widehat{H}_n - H)^2\})^{1/2} = O\left(n^{-1/2}\log n\right).$$

The following corollary is an easy consequence of Theorem 7. In particular, part (b) follows along the same lines as the proof of Theorem 7, leading to equation (11) below.

Corollary 5 *(a) The plug-in estimates can tend to H at an arbitrarily slow algebraic rate $O(n^{-\epsilon})$ even when $H^{(r)} < \infty$ for* all *$r \geq 1$.*

(b) Assume that for some $q > 2$ there exist positive constants c_1, $c_2 > 0$ such that $\frac{c_1}{i \log^q i} \leq p(i) \leq \frac{c_2}{i \log^q i}$, $i = 1, 2, \ldots$. Then,

$$H - \mathbf{E}(\widehat{H}_n) = \Omega(1/\log^{q-1} n),$$

so the convergence to H is even slower, despite the fact that $H^{(q-1-\epsilon)}$ is finite.

Proof of Theorem 7. By (5), and properties (ii) and (iv) from Section 2, it suffices to prove that

$$H - \mathbf{E}(\widehat{H}_n) = \Theta\left(n^{-\frac{q-1}{q}}\right).$$

First we show that

$$H - \mathbf{E}(\widehat{H}_n) \leq \sum_{i=1}^{\infty} p(i) \log_2 \left(1 + \frac{1 - p(i)}{np(i)}\right).$$

Observe that if np_n is a Binomial(n, p) random variable and $(n-1)p_{n-1}$ is a Binomial$(n-1, p)$ random variable, then

$$\begin{aligned}
\mathbf{E}\{-p_n \log_2 p_n\} &= -\sum_{k=0}^{n} \frac{k}{n} \log_2 \frac{k}{n} \binom{n}{k} p^k (1-p)^{n-k} \\
&= -p \sum_{k=1}^{n} \log_2 \frac{k}{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} \\
&= -p \sum_{k=0}^{n-1} \log_2 \frac{k+1}{n} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\
&= -p\mathbf{E}\left\{\log_2 \left(\frac{(n-1)p_{n-1} + 1}{n}\right)\right\} \geq -p \log_2 \left(\frac{(n-1)p + 1}{n}\right),
\end{aligned}$$

where in the last step we applied Jensen's inequality for the concave function $\log_2 x$. Applying this for every $p_n(i)$ and summing over $i \geq 1$ gives

$$\mathbf{E}(\widehat{H}_n) \geq -\sum_{i=1}^{\infty} p(i) \log_2 \left(\frac{(n-1)p(i) + 1}{n}\right),$$

and thus

$$H - \mathbf{E}(\widehat{H}_n) \leq \sum_{i=1}^{\infty} p(i) \log_2 \left(1 + \frac{1 - p(i)}{np(i)}\right) \leq \sum_{i=1}^{\infty} p(i) \log_2 \left(1 + \frac{1}{np(i)}\right).$$

Splitting this sum into two terms

$$\sum_{i=1}^{\infty} p(i) \log_2 \left(1 + \frac{1}{np(i)}\right) \leq \sum_{i: np(i) \leq 1} p(i) \log_2 \frac{2}{np(i)} + \sum_{i: np(i) > 1} \frac{1}{n}.$$

Now taking the bounds on $p(i)$, $p(i) > 1/n$ implies $i < (c_2 n)^{1/q}$ and $p(i) \le 1/n$ implies $i \ge (c_1 n)^{1/q}$. So for the second term

$$\sum_{i:np(i)>1} \frac{1}{n} \le \frac{\lfloor (c_2 n)^{\frac{1}{q}} \rfloor}{n} \le c_2^{\frac{1}{q}} n^{-\frac{q-1}{q}}.$$

For n large enough, the first term

$$\begin{aligned} \sum_{i:np(i)\le 1} p(i) \log_2 \frac{2}{np(i)} \quad &\le \quad \sum_{i \ge (c_1 n)^{1/q}} \frac{c_2}{i^q} \log_2 \frac{2i^q}{nc_1} \quad \le \quad O\left(\frac{1}{n}\right) + \int\limits_{(c_1 n)^{1/q}}^{\infty} \frac{c_2}{x^q} \log_2 \frac{ex^q}{nc_1}\, dx \\ &= \quad O\left(\frac{1}{n}\right) + \frac{c_2}{n^{1-1/q} q} \int\limits_{c_1}^{\infty} \frac{1}{u^{2-1/q}} \log_2 \frac{eu}{c_1}\, du \quad = \quad O\left(n^{-\frac{q-1}{q}}\right), \end{aligned}$$

which gives the upper bound.

For the lower bound, recall that property (ii) in Section 2 was obtained from the fact that

$$\mathbf{E}\{p_n(i) \log_2 p_n(i)\} \ge p(i) \log_2 p(i) \quad \text{for all } i \text{ (Jensen's inequality).}$$

Thus

$$\begin{aligned} H - \mathbf{E}(\widehat{H}_n) \quad &= \quad \sum_{i=1}^{\infty} (\mathbf{E}\{p_n(i) \log_2 p_n(i)\} - p(i) \log_2 p(i)) \\ &\ge \quad \sum_{i:np(i)\le 1/2} (\mathbf{E}\{p_n(i) \log_2 p_n(i)\} - p(i) \log_2 p(i)) \\ &= \quad \frac{1}{n} \sum_{i:np(i)\le 1/2} (\mathbf{E}\{np_n(i) \log_2 np_n(i)\} - np(i) \log_2 np(i)) \\ &\ge \quad - \sum_{i:np(i)\le 1/2} p(i) \log_2 np(i) \quad \ge \quad \sum_{i:np(i)\le 1/2} p(i), \end{aligned} \tag{11}$$

because $np_n(i) \log_2 np_n(i)$ is always nonnegative. By the given bounds,

$$H - \mathbf{E}(\widehat{H}_n) \ge \sum_{i \ge (2c_2 n)^{1/q}} c_1 i^{-q} \ge c_1 \int\limits_{\lceil (2c_2 n)^{1/q} \rceil}^{\infty} \frac{1}{x^q}\, dx = \frac{c_1}{q-1} \frac{1}{\lceil (2c_2 n)^{1/q} \rceil^{q-1}} = \Omega\left(n^{-\frac{q-1}{q}}\right),$$

which concludes the proof. $\square$

In the case of mutual information, using the identity $I = H(V) + H(W) - H(V,W)$, the results for entropy estimation imply that the same upper bound of order $n^{-\frac{q-1}{q}}$ for $q < 2$ holds for the error of $\{\widehat{I}_n\}$:

Corollary 6 *Assume that the tail condition of Theorem 7 holds for the distributions of V and W, and also for their joint distribution. Then, for the plug-in estimate,*

$$\mathbf{E}\{|\widehat{I}_n - I|\} \le (\mathbf{E}\{(\widehat{I}_n - I)^2\})^{1/2} = \begin{cases} O\left(n^{-\frac{q-1}{q}}\right) & \text{if } q < 2, \\ O\left(n^{-1/2} \log n\right) & \text{if } q \ge 2. \end{cases}$$

5 Entropy Estimation: Match-Lengths

In this subsection we provide convergence rates for a different entropy estimator, based on match-lengths. This approach is inspired by the Lempel-Ziv family of data compression algorithms [28][29], and has been very successful in nonparametric entropy estimation from processes with memory (see [22][18] and the references therein). Here we use the simplest form of a match-length entropy estimator, to demonstrate that it is possible to obtain convergence rates under the weaker (and, in view of the above heuristic, also somewhat more natural) assumption that $H^{(r)} = \mathbf{E}\{(-\log_2 p(X))^r\}$ is finite for some $r \geq 2$.

Given a sample $x_1^n = (x_1, x_2, \ldots, x_n)$ of the i.i.d. random variables $(X_1, X_2, \ldots, X_n)$, we write $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$ for $1 \leq i \leq j \leq n$. For any $n \geq 1$ we define the *match-length* L_n as the length L of the shortest initial prefix x_1^L that does not match anywhere else in x_1^n

$$L_n = \min\{1 \leq L \leq n \ : \ x_1^L \neq x_{j+1}^{j+L} \ \text{ for all } \ 1 \leq j \leq n-L\},$$

with the convention that the minimum of the empty set equals n. Alternatively, L_n can be thought of as the length of the longest matching prefix, plus one.

For example, if $x_1^n = abbbcbabbaac$ (with $n = 12$) then $L_n = 4$ since abb appears twice in x_1^n but $abbb$ appears only once. Also, if x_1^n is a constant sequence of the form $aaa \cdots a$, then $L_n = n$ by convention.

Based on the match-lengths L_n, for each $n \geq 1$ we define the following entropy estimators:

$$\widetilde{H}_n = \frac{\log_2 n}{L_n}.$$

Below we will prove the following results about the $\widetilde{H}_n$:

Theorem 8 (a) Consistency. *If $H < \infty$, then*

$$\lim_{n\to\infty} \widetilde{H}_n = H \quad a.s.$$

(b) Convergence rate in probability. *If $H^{(2)} = \mathbf{E}\{(-\log_2 p(X))^2\} < \infty$, then, as $n \to \infty$,*

$$\sqrt{\log_2 n}\left(\widetilde{H}_n - H\right) \to N(0, H\sigma^2) \quad \textit{in distribution,}$$

where $\sigma^2 = \mathbf{Var}\{-\log_2 p(X)\}$. In particular,

$$\widetilde{H}_n \ = \ H + O_P\left(\frac{1}{\sqrt{\log n}}\right).$$

(c) L^1 and L^2 lower bounds. *If $\sigma^2 \neq 0$ and $H^{(2)} < \infty$, then*

$$\mathbf{E}\{|\widetilde{H}_n - H|\} = \Omega\left(\frac{1}{\sqrt{\log n}}\right)$$

and

$$\mathbf{E}\{(\widetilde{H}_n - H)^2\} = \Omega\left(\frac{1}{\log n}\right).$$

Recall that for a sequence of random variables $\{Y_n\}$ and a sequence of nonnegative real numbers $\{a_n\}$ we say $Y_n = O_P(a_n)$ if and only if the sequence of distributions of the random variables $\{Y_n/a_n\}$ is tight.

Note that, although the discussion here is restricted entirely to the case of i.i.d. random variables, the estimator $\widetilde{H}_n$ is consistent for arbitrary stationary and ergodic sources; cf. [26][17].

Our next result says that under the stronger assumption that $H^{(4)} < \infty$, the above lower bound on the L^2 error of $\widetilde{H}_n$ is tight.

Theorem 9 *If* $H^{(4)} = \mathbf{E}\{(-\log_2 p(X))^4\} < \infty$, *then*

$$\mathbf{E}\{(\widetilde{H}_n - H)^2\} = O\left(\frac{1}{\log n}\right).$$

Following [26], in order to analyze the asymptotics of L_n we introduce the *recurrence times* R_m, where, for $m \geq 1$, R_m denotes the time of the first recurrence of the initial m-block x_1^m in the realization $x_1, x_2, \ldots$:

$$R_m = \inf\{k > m \ : \ x_1^m = x_{k-m+1}^k\}.$$

Observe that R_m and L_n are related via the following duality relationship:

$$R_m \leq n \quad \text{iff} \quad L_n \geq m + 1. \tag{12}$$

The proof of Theorem 8 follows along the lines of the corresponding match-length results in [17], but since here the alphabet $\mathcal{X}$ is infinite, most of the combinatorial arguments need to be modified.

Before turning to the proofs, note that the case $H = 0$ is trivial: When $H = 0$ the random variable X takes on only one value and, by the definition of L_n, $\widetilde{H}_n = (\log_2 n)/n$ for all n. In this case the results in Theorems 8 and 9 all trivially hold. Therefore, from now on we assume, without loss of generality, that $H \neq 0$. The following lemma will be used repeatedly in the proofs of Theorems 8 and 9; its proof is in the Appendix.

Lemma 3 (i) *There is a finite constant C such that, for all $n \geq 1$, any string x_1^n of nonzero probability $p^n(x_1^n)$, and any $\epsilon > 0$,*

$$\mathbf{P}\{\log_2[R_n p^n(x_1^n)] \geq \epsilon\sqrt{n} \,|\, x_1^n\} \leq C 2^{-\epsilon\sqrt{n}}.$$

(ii) *For all $n \geq 1$, any string x_1^n of nonzero probability $p^n(x_1^n)$, and any $\epsilon > 0$,*

$$\mathbf{P}\{\log_2[2np^n(x_1^n)] \leq \log_2[R_n p^n(x_1^n)] \leq -\epsilon\sqrt{n} \,|\, x_1^n\} \leq 2^{-\epsilon\sqrt{n}}.$$

(iii) *There are finite constants $\alpha, \beta > 0$ such that, for all $n \geq 1$,*

$$\mathbf{P}\{\log_2[R_n p^n(X_1^n)] < \log_2[2np^n(X_1^n)]\} = \mathbf{P}\{n + 1 \leq R_n \leq 2n - 1\} \leq \alpha 2^{-\beta n}.$$

Proof of Theorem 8. Since the sequence $\{2^{-\epsilon\sqrt{m}}\}$ is summable over m for any $\epsilon > 0$, Lemma 3 together with the Borel-Cantelli lemma implies that

$$\frac{1}{\sqrt{m}} \log_2[R_m p^m(X_1^m)] \to 0 \ \text{ a.s.} \quad \text{ as } \ m \to \infty. \tag{13}$$

In particular, by an application of the strong law of large numbers,

$$\lim_{m\to\infty} \frac{1}{m} \log_2 R_m = \lim_{m\to\infty} -\frac{1}{m} \log_2 p^m(X_1^m) = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} [-\log_2 p(X_i)] = H \ \text{ a.s.}$$

This together with the duality relationship (12) implies that

$$\frac{L_n}{\log_2 n} \to \frac{1}{H} \ \text{ a.s.} \tag{14}$$

and this proves part (a). Similarly, (13) can be rewritten as

$$\frac{1}{\sqrt{m}}[\log_2 R_m - mH] - \frac{1}{\sqrt{m}}[-\log_2 p^m(X_1^m) - mH] \to 0 \ \text{ a.s.}$$

But the second term, by the central limit theorem, is asymptotically Normal with mean zero and variance $\sigma^2 = \mathbf{Var}\{-\log_2 p(X)\}$, therefore

$$\frac{1}{\sqrt{m}}[\log_2 R_m - mH] \to N(0, \sigma^2) \quad \text{in distribution, as } \ m \to \infty.$$

This, together with the duality relationship (12), implies that

$$\frac{1}{\sqrt{\log_2 n}} \left[L_n - \frac{\log_2 n}{H} \right] \to N(0, H^{-3}\sigma^2) \quad \text{in distribution,}$$

and combining this with (14) easily shows that

$$\sqrt{\log_2 n} \left(\widetilde{H}_n - H \right) \to N(0, H\sigma^2) \quad \text{in distribution.} \tag{15}$$

The convergence rate in probability follows immediately.

For part (c) note that, for $K > 0$, by (15),

$$\mathbf{P}\left\{ \sqrt{\log_2 n} |\widetilde{H}_n - H| > K \right\} \to 2 - 2\Phi\left(\frac{K}{\sigma\sqrt{H}} \right)$$

where $\Phi(x)$ denotes the standard Gaussian distribution function. Therefore,

$$\begin{aligned} \mathbf{E}\left\{ \sqrt{\log_2 n} |\widetilde{H}_n - H| \right\} &\geq K\mathbf{P}\left\{ \sqrt{\log_2 n} |\widetilde{H}_n - H| > K \right\} \\ &\to 2K \left[1 - \Phi\left(\frac{K}{\sigma\sqrt{H}} \right) \right] > 0, \end{aligned}$$

i.e., $\mathbf{E}\{|\widetilde{H}_n - H|\} = \Omega(1/\sqrt{\log n})$. This also implies that

$$\mathbf{E}\{(\widetilde{H}_n - H)^2\} \geq [\mathbf{E}\{|\widetilde{H}_n - H|\}]^2 = \Omega(1/\log n)$$

and completes the proof. $\square$

Proof of Theorem 9. It suffices to show that for the random variables

$$Z_n = \sqrt{\log_2 n}\left[\tilde{H}_n - H\right] = \sqrt{\log_2 n}\left[\frac{\log_2 n}{L_n} - H\right]$$

the sequence $\mathbf{E}\{Z_n^2\}$ is bounded in n. Write ρ_4 for the (centralized) fourth moment

$$\rho_4 = \mathbf{E}\{(-\log_2 p(X) - H)^4\}.$$

We expand

$$\begin{aligned}
\mathbf{E}\{Z_n^2\} &= \int_0^\infty \mathbf{P}\{Z_n^2 \geq x\}dx \\
&\leq 2 + \int_2^{\log_2 n(\log_2 n - H)^2} \mathbf{P}\left\{L_n \leq \frac{\log_2 n}{H + \sqrt{\frac{x}{\log_2 n}}}\right\} dx \\
&\quad + \int_2^{\log_2 n(H-(\log_2 n)/n)^2} \mathbf{P}\left\{L_n \geq \frac{\log_2 n}{H - \sqrt{\frac{x}{\log_2 n}}}\right\} dx,
\end{aligned} \tag{16}$$

and treat the two integrals above separately.

For the first term, let

$$m = m(n,x) \triangleq \left\lfloor \frac{\log_2 n}{H + \sqrt{\frac{x}{\log_2 n}}} \right\rfloor$$

and note that by the duality relationship (12), the probability $\mathbf{P}\{L_n \leq m\}$ is equal to

$$\begin{aligned}
&\mathbf{P}\{R_m > n\} \\
&= \sum_{x_1^m \in \mathcal{X}^m} p^m(x_1^m)\mathbf{P}\left\{ \frac{1}{\sqrt{m}}\log_2[p^m(x_1^m)R_m] > \frac{1}{\sqrt{m}}\log_2[p^m(x_1^m)n] \,\middle|\, x_1^m \right\} \\
&\leq \sum_{x_1^m : p^m(x_1^m) > \gamma_n} p^m(x_1^m)\mathbf{P}\left\{ \frac{1}{\sqrt{m}}\log_2[p^m(x_1^m)R_m] > \frac{1}{\sqrt{m}}\log_2[p^m(x_1^m)n] \,\middle|\, x_1^m \right\} \\
&\quad + \sum_{x_1^m : p^m(x_1^m) \leq \gamma_n} p^m(x_1^m)
\end{aligned}$$

where $\gamma_n = (\log_2 n)^3/n$. By Lemma 3 (i), this is bounded above by

$$\begin{aligned}
&\sum_{x_1^m : p^m(x_1^m) > \gamma_n} C/n \; + \; \mathbf{P}\{p^m(X_1^m) \leq \gamma_n\} \\
&\leq \frac{C}{n}\left|\{x_1^m : p^m(x_1^m) > \gamma_n\}\right| + \mathbf{P}\{-\log_2 p^m(X_1^m) - mH \geq \log_2 n - 3\log_2\log_2 n - mH\}.
\end{aligned}$$

Using Markov's inequality and the fact that there cannot be more than $1/\gamma_n$ strings with probability greater than γ_n, this is bounded above by

$$\begin{aligned}
\frac{C}{n}\frac{1}{\gamma_n} \quad + \quad & \frac{m\rho_4 + 3m(m-1)\sigma^4}{(\log_2 n - 3\log_2\log_2 n - mH)^4} \\
\leq \quad & \frac{C}{(\log_2 n)^3} + \frac{C' m^2}{(\log_2 n - 3\log_2\log_2 n - mH)^4} \\
\leq \quad & \frac{C}{(\log_2 n)^3} + C' \frac{m^2/(\log_2 n)^4}{\left(1 - \frac{H}{H+\sqrt{x/\log_2 n}} - \frac{3\log_2\log_2 n}{\log_2 n}\right)^4} \\
\leq \quad & \frac{C}{(\log_2 n)^3} + C' \frac{\left(\frac{1}{H+\sqrt{x/\log_2 n}}\right)^2 \left(\frac{1}{\log_2 n}\right)^2}{\left[\frac{1}{2}\left(1 - \frac{H}{H+\sqrt{x/\log_2 n}}\right)\right]^4}
\end{aligned}$$

where the last inequality follows from the definition of m, and the observation that, for all $n \geq$ some N_0 (independent of x),

$$3\frac{\log_2\log_2 n}{\log_2 n} \leq \frac{1}{2}\left(1 - \frac{H}{H + \sqrt{x/\log_2 n}}\right),$$

for all $x \geq 2$. Simplifying the above expression, we have shown that

$$\mathbf{P}\{L_n \leq m\} \leq \frac{C}{(\log_2 n)^3} + \frac{C''}{x^2}\left(H + \sqrt{\frac{x}{\log_2 n}}\right)^2.$$

Therefore, for $n \geq N_0$ the first integral term in (16) is bounded above by

$$\int_2^{(\log_2 n)^3} \frac{C\,dx}{(\log_2 n)^3} \; + \; C'' \int_2^{(\log_2 n)^3} \frac{1}{x^2}\left(H + \sqrt{\frac{x}{\log_2 n}}\right)^2 dx$$

and simply evaluating these two integrals shows that they are bounded in n, and hence

$$\int_2^{\log_2 n(\log_2 n - H)^2} \mathbf{P}\{L_n \leq m\} dx \;=\; O(1). \tag{17}$$

For the second integral term in (16) write

$$M = M(n,x) \triangleq \left\lceil \frac{\log_2 n}{H - \sqrt{\frac{x}{\log_2 n}}} \right\rceil$$

and let $N = M - 1$. By the duality relationship (12), the probability $\mathbf{P}\{L_n \geq M\}$ is equal to

$$\begin{aligned}
&\mathbf{P}\{R_N \leq n\} \\
&\quad = \; \mathbf{P}\{M \leq R_N \leq 2M-3\} + \mathbf{P}\{2M-2 \leq R_N \leq n\} \\
&\quad \leq \; \alpha 2^{-\beta N} + \sum_{x_1^N \in \mathcal{X}^N} p^N(x_1^N)\mathbf{P}\{\log_2[2Np^N(x_1^N)] \leq \log_2[R_N p^N(x_1^N)] \leq \log_2[np^N(x_1^N)] \,\big|\, x_1^N\}
\end{aligned}$$

where the inequality follows from Lemma 3 (iii). Letting $\delta_n = (n \log_2 n)^{-1}$ and using Lemma 3 (ii), we obtain that

$$\begin{aligned}
\mathbf{P}\{L_n \geq M\} &\leq \alpha 2^{-\beta N} + \sum_{x_1^N : p^N(x_1^N) > \delta_n} p^N(x_1^N) + \sum_{x_1^N : p^N(x_1^N) \leq \delta_n} n[p^N(x_1^N)]^2 \\
&\leq \alpha' 2^{-\beta M} + \mathbf{P}\{p^N(X_1^N) > \delta_n\} + \sum_{x_1^N : p^N(x_1^N) \leq \delta_n} p^N(x_1^N) n \delta_n \\
&\leq \alpha' 2^{-\beta M} + \mathbf{P}\{p^N(X_1^N) > \delta_n\} + \frac{1}{\log_2 n}. \qquad (18)
\end{aligned}$$

Now we claim (to be verified below) that for $n \geq$ some N_1, uniformly in $x \geq 2$, we have

$$\mathbf{P}\{p^N(X_1^N) > \delta_n\} \leq C \left(\frac{\log_2 n}{Mx} \right)^2 \qquad (19)$$

(with C denoting an absolute constant, not the same as in the derivation of the previous part). From (18) and (19) we then get that, for n large enough,

$$\begin{aligned}
&\int_2^{\log_2 n (H - (\log_2 n)/n)^2} \mathbf{P}\{L_n \geq M\} dx \\
&\quad \leq \int_2^{H^2 \log_2 n} \alpha' 2^{-\beta M(n,x)} dx + C \int_2^{H^2 \log_2 n} \left(\frac{\log_2 n}{M} \right)^2 \frac{1}{x^2} dx + H^2 \\
&\quad \leq H^2 \alpha' (\log_2 n) 2^{-\beta M(n,0)} + C \int_2^{H^2 \log_2 n} \frac{1}{x^2} \left(H - \sqrt{\frac{x}{\log_2 n}} \right)^2 dx + H^2 \\
&\quad \leq \frac{H^2 \alpha' \log_2 n}{n^{\beta/H}} + O(1) + H^2
\end{aligned}$$

where the last inequality is easily verified by evaluating the above integral explicitly. Therefore,

$$\int_2^{\log_2 n (H - (\log_2 n)/n)^2} \mathbf{P}\{L_n \geq M\} dx \;=\; O(1),$$

and combining this with (17) and (16) implies that $\mathbf{E}\{Z_n^2\} = O(1)$ and completes the proof of the Theorem.

Finally it remains to establish the claim (19). Using Markov's inequality and the fourth moment assumption as before,

$$\begin{aligned}
\mathbf{P}\{p^N(X_1^N) > \delta_n\} &= \mathbf{P}\{\log_2 p^N(X_1^N) + NH > \log_2 \delta_n + NH\} \\
&\leq \frac{N\rho_4 + 3N(N-1)\sigma^4}{(\log_2 n + \log_2 \log_2 n - NH)^4} \\
&\leq \frac{CM^2}{(\log_2 n + \log_2 \log_2 n - MH + H)^4} \\
&\leq C \left(\frac{\log_2 n}{Mx} \right)^2 \left[\frac{1}{M} \left(\frac{\log_2 n}{x} \right)^{1/2} ((\log_2 \log_2 n + H) - (MH - \log_2 n)) \right]^{-4}
\end{aligned}$$

so it suffices to show that the term $[\cdots]^{-4}$ above is uniformly bounded over $x \geq 2$. By the definition of M, $M(n,x) \geq M(n,2)$, and it is easy to see that for n large enough $MH - \log_2 n \geq \frac{\sqrt{2}}{H}\sqrt{\log_2 n}$. Therefore, for all $n \geq$ some N_1 (independent of x),

$$\log_2 \log_2 n + H \leq \frac{1}{H\sqrt{2}}\sqrt{\log_2 n} \leq \frac{1}{2}(MH - \log_2 n).$$

Noting that for n large enough (uniformly in $x \geq 2$)

$$M \leq \frac{2\log_2 n}{H - \sqrt{\frac{x}{\log_2 n}}}$$

and substituting the last two bounds in the expression $[\cdots]^{-4}$ above, we get

$$[\cdots]^{-4} \leq \left[\frac{2M\sqrt{x}}{\sqrt{\log_2 n}(MH - \log_2 n)}\right]^4 \leq 256\left[\frac{\frac{\log_2 n}{H - \sqrt{\frac{x}{\log_2 n}}}\sqrt{x}}{\sqrt{\log_2 n}\left(H\frac{\log_2 n}{H - \sqrt{\frac{x}{\log_2 n}}} - \log_2 n\right)}\right]^4 = 256,$$

as required. $\square$

Acknowledgements. We wish to thank László Györfi and Gábor Lugosi for stimulating and helping this work. We also thank the associate editor Wojtek Szpankowski and the two anonymous reviewers for helping improve the quality of the presentation of our results.

Appendix

In the notation of Section 2, we state and prove the four lemmas that were used in the proofs of Theorems 1 and 2.

Lemma 4 *Assume that g is concave and monotone increasing, and that (4) is satisfied. Then*

$$\limsup_{n\to\infty} \mathbf{E}(\widehat{F}_n) \le F.$$

Remark. Note that the assumption $f \ge 0$ can be replaced by the condition that $\sum_i \|\min(f_i, 0)\|_\infty < \infty$, and that in this case the above lemma as well as the following two lemmas still hold. Similarly, if the assumption that $f \ge 0$ is replaced by the condition that $\sum_i \|\max(f_i, 0)\|_\infty < \infty$, then the lemmas remain valid after interchanging the terms 'convex' and 'concave,' the terms $\le$ and $\ge$, and the terms '$\liminf$' and '$\limsup$.'

Lemma 5 *Assume that all the f_i are continuous, and that g is monotone increasing and continuous. Then*

$$\liminf_{n\to\infty} \widehat{F}_n \ge F \quad a.s. \qquad \text{and} \qquad \liminf_{n\to\infty} \mathbf{E}(\widehat{F}_n) \ge F.$$

Lemma 6 *(a) If g is concave and monotone increasing, and all the f_i are continuous and they satisfy (4), then $\{\widehat{F}_n\}$ is asymptotically unbiased, that is,*

$$\lim_{n\to\infty} \mathbf{E}(\widehat{F}_n) = F.$$

(b) If g is linear and both f and $-f$ satisfy (4), that is, $\lim_{n\to\infty} \sum_i \mathbf{E}[f_i(p_n(i))] = \sum_i f_i(p(i))$, then $\{\widehat{F}_n\}$ is asymptotically unbiased. If, moreover, g and all the f_i are linear, then $\widehat{F}_n$ is unbiased, that is, $\mathbf{E}(\widehat{F}_n) = F$.

Lemma 7 *(Here f is not necessarily nonnegative.) Assume that g and all the f_i are continuous, and that $L = \sum_i \|f_i\|_\infty < \infty$. Then $\{\widehat{F}_n\}$ is strongly universally consistent, and also L^p consistent for any $p > 0$, that is,*

$$\lim_{n\to\infty} \widehat{F}_n = F \quad a.s. \qquad \text{and} \qquad \lim_{n\to\infty} \mathbf{E}\{|\widehat{F}_n - F|^p\} = 0.$$

In particular, these hold if g and all the f_i are continuous, and the support of X is finite.

Proof of Lemma 4. Note that if g is concave on $\mathcal{R}$ then it is also continuous. Using Jensen's inequality, the concavity, continuity, and monotonicity of g, and (4), we get

$$\limsup_{n\to\infty} \mathbf{E}(\widehat{F}_n) \le \limsup_{n\to\infty} g\left(\sum_i \mathbf{E}[f_i(p_n(i))]\right) \le g\left(\limsup_{n\to\infty} \sum_i \mathbf{E}[f_i(p_n(i))]\right) \le g\left(\sum_i f_i(p(i))\right) = F. \quad \square$$

Note that the proof essentially remains the same if instead of $f \ge 0$ we assume $f \le 0$. Therefore, if g is convex and monotone increasing, and $-f$ satisfies (4), then $\liminf_{n\to\infty} \mathbf{E}(\widehat{F}_n) \ge F$.

PROOF OF LEMMA 5. Noting that, by the strong law of large numbers, for each i, $\lim_{n\to\infty} p_n(i) = p(i)$ a.s., continuity and Fatou's lemma imply

$$\liminf_{n\to\infty} \sum_i f_i(p_n(i)) \geq \sum_i \liminf_{n\to\infty} f_i(p_n(i)) = \sum_i f_i(p(i)) \quad \text{a.s.}$$

Since g is monotone increasing and continuous on $(0,\infty)$

$$\liminf_{n\to\infty} \widehat{F}_n = \liminf_{n\to\infty} g\left(\sum_i f_i(p_n(i))\right) \geq g\left(\liminf_{n\to\infty} \sum_i f_i(p_n(i))\right) \geq g\left(\sum_i f_i(p(i))\right) = F \quad \text{a.s.}$$

Finally, by the monotonicity of g, $\widehat{F}_n \geq g(0)$ for all n, hence by Fatou's lemma,

$$\liminf_{n\to\infty} \mathbf{E}(\widehat{F}_n) \geq \mathbf{E}\{\liminf_{n\to\infty} \widehat{F}_n\} \geq F. \quad \Box$$

PROOF OF LEMMA 6. Obvious from Lemma 4 and 5. □

PROOF OF LEMMA 7. Noting that for each i, $\lim_{n\to\infty} p_n(i) = p(i)$ a.s., continuity and the dominated convergence theorem (using $L < \infty$) imply

$$\lim_{n\to\infty} \sum_i f_i(p_n(i)) = \sum_i \lim_{n\to\infty} f_i(p_n(i)) = \sum_i f_i(p(i)) \quad \text{a.s.}$$

Since g is continuous

$$\lim_{n\to\infty} \widehat{F}_n = \lim_{n\to\infty} g\left(\sum_i f_i(p_n(i))\right) = g\left(\lim_{n\to\infty} \sum_i f_i(p_n(i))\right) = g\left(\sum_i f_i(p(i))\right) = F \quad \text{a.s.}$$

Observing that $|F|$ and the functions $|\widehat{F}_n|$ are all bounded (all taking values in the bounded set $g([-L, L])$), the consistency in L^p follows from the dominated convergence theorem.

The continuity of all the f_i and the finiteness of the support imply $L < \infty$. □

Next we give the proofs of four results that were stated and used without a proof earlier.

PROOF OF COROLLARY 3. Observe that $n^a \widehat{F}_n^{(a)} = \sum_i (np_n(i))^a$ satisfies Proposition 3 with

$$G^{(a)}(x_1, \ldots, x_{n-1}) = \sum_i \left(\sum_{j=1}^{n-1} I_{\{x_j = i\}}\right)^a,$$

because

$$\begin{aligned}
0 &< n^a \widehat{F}_n^{(a)}(x_1, \ldots, x_n) - G^{(a)}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \\
&= \sum_i \left((np_n(i))^a - (np_n(i) - I_{\{x_j=i\}})^a\right) = (np_n(x_j))^a - (np_n(x_j) - 1)^a \\
&= (np_n(x_j))^a \left(1 - \left(1 - \frac{1}{np_n(x_j)}\right)^a\right) \leq (np_n(x_j))^a \left(1 - \left(1 - \frac{1}{np_n(x_j)}\right)\right) \\
&= (np_n(x_j))^{a-1} \leq 1,
\end{aligned}$$

and so

$$\sum_{j=1}^{n}\left(n^a\widehat{F}_n^{(a)}(x_1,\ldots,x_n)-G^{(a)}(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)\right)$$
$$\leq\sum_{j=1}^{n}(np_n(x_j))^{a-1}=\sum_i(np_n(i))^a=n^a\widehat{F}_n^{(a)}.$$

This implies that, for any $\epsilon>0$,

$$\mathbf{P}\{|\widehat{F}_n^{(a)}-\mathbf{E}[\widehat{F}_n^{(a)}]|\geq\epsilon\}=\mathbf{P}\{|n^a\widehat{F}_n^{(a)}-n^a\mathbf{E}[\widehat{F}_n^{(a)}]|\geq n^a\epsilon\}\leq 2e^{-n^a\epsilon^2/(2\mathbf{E}[\widehat{F}_n^{(a)}]+2\epsilon/3)}.\ \square$$

Proof of Lemma 1. It is well-known (see, e.g., [21, p.137]) that the a^{th} moment of a binomial random variable can be expressed as

$$\mathbf{E}\{\mathrm{Bin}^a(n,p)\}=\sum_{k=1}^{a-1}\sigma_a^{(k)}\frac{n!}{(n-k)!}p^k+\frac{n!}{(n-a)!}p^a\leq\sum_{k=1}^{a-1}\sigma_a^{(k)}n^kp^k+n^ap^a$$

where the $\sigma_a^{(k)}$ are Stirling numbers of the second kind. This implies that

$$\sum_i\mathbf{E}\{p_n^a(i)\}=\sum_i\mathbf{E}\{\mathrm{Bin}^a(n,p(i))/n^a\}\leq\sum_{k=1}^{a-1}\sum_i p^k(i)\sigma_a^{(k)}n^{k-a}+\sum_i p^a(i)=O(1/n)+\sum_i p^a(i).\ \square$$

Proof of Lemma 2. We use randomization such that u is replaced by a sequence $U=(U_0,U_1,U_2,\ldots)$ of i.i.d. Bernoulli(1/2) random variables, independent of the samples $(X_1,\ldots,X_n)$. Let U_+^k and U_-^k denote the sequence U with the difference that U_+^k forces the k^{th} bit to be 1 and U_-^k forces the k^{th} bit to be 0. Introduce the notation

$$r_n(u)=\mathbf{P}\{|F_n(D_n(u))-F(u)|>a_n\}.$$

Now for $n,k\in\mathcal{N}$

$$\begin{aligned}\mathbf{E}\{r_n(U)\} &= \mathbf{E}\{\mathbf{P}\{|F_n(D_n(U))-F(U)|>a_n|U\}\}\\ &= \mathbf{P}\{|F_n(D_n(U))-F(U)|>a_n\}\\ &\geq \mathbf{E}\{\mathbf{P}\{|F_n(D_n(U))-F(U)|>a_n|D_n(U)\}I_{\{D_n(U)\in A_{n,k}\}}\}.\end{aligned}$$

Conditioning on $D_n(U)$ in $A_{n,k}$,

$$\begin{aligned}&\mathbf{P}\{|F_n(D_n(U))-F(U)|>a_n|D_n(U)\}\\ &\quad= \mathbf{P}\{|F_n(D_n(U))-F(U_-^k)|>a_n,U_k=0|D_n(U)\}\\ &\qquad+\mathbf{P}\{|F_n(D_n(U))-F(U_+^k)|>a_n,U_k=1|D_n(U)\}\\ &\quad= \mathbf{P}\{|F_n(D_n(U))-F(U_-^k)|>a_n|D_n(U)\}\mathbf{P}\{U_k=0|D_n(U)\}\\ &\qquad+\mathbf{P}\{|F_n(D_n(U))-F(U_+^k)|>a_n|D_n(U)\}\mathbf{P}\{U_k=1|D_n(U)\},\end{aligned}$$

because (9) implies that $\mathbf{P}\{U_k = 1 | U_-^k, D_n(U)\} = 1/2$ for every possible U_-^k and $D_n(U) \in A_{n,k}$, and thus U_-^k and U_k (or similarly U_+^k and U_k) are conditionally independent given $D_n(U) \in A_{n,k}$. Moreover, $\mathbf{P}\{U_k = 1 | D_n(U)\} = 1/2$ for $D_n(U) \in A_{n,k}$, so

$$\begin{aligned}
&\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| > a_n | D_n(U)\}\mathbf{P}\{U_k = 0 | D_n(U)\} \\
&\qquad + \mathbf{P}\{|F_n(D_n(U)) - F(U_+^k)| > a_n | D_n(U)\}\mathbf{P}\{U_k = 1 | D_n(U)\} \\
&= \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| > a_n | D_n(U)\} + \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_+^k)| > a_n | D_n(U)\} \\
&\geq \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| + |F_n(D_n(U)) - F(U_+^k)| > 2a_n | D_n(U)\} \\
&\geq \frac{1}{2}\mathbf{P}\{|F(U_-^k) - F(U_+^k)| > 2a_n | D_n(U)\} \\
&\geq \frac{I_{\{d_k > 2a_n\}}}{2} \qquad \text{(using (10)).}
\end{aligned}$$

Therefore, taking the expectation of this chain of inequalities on $A_{n,k}$, we have for any $n, k \in \mathcal{N}$

$$\mathbf{E}\{r_n(U)\} \geq \frac{I_{\{d_k > 2a_n\}}}{2}\mathbf{P}\{D_n(U) \in A_{n,k}\} \geq \frac{I_{\{d_k > 2a_n\}}}{2} \inf_u \mu_u^{(n)}(A_{n,k}),$$

and in particular, for $k \in \mathcal{N}$ and $n = n_k$

$$\mathbf{E}\{r_{n_k}(U)\} \geq \frac{I_{\{d_k > 2a_{n_k}\}}}{2} \inf_u \mu_u^{(n_k)}(A_{n_k,k}).$$

Since $r_n(U) \leq 1$, by Fatou's lemma

$$\begin{aligned}
\sup_u \limsup_{n\to\infty} r_n(u) &\geq \mathbf{E}\{\limsup_{n\to\infty} r_n(U)\} \\
&\geq \limsup_{n\to\infty} \mathbf{E}\{r_n(U)\} \\
&\geq \limsup_{k\to\infty} \mathbf{E}\{r_{n_k}(U)\} \\
&\geq \frac{1}{2}\limsup_{k\to\infty}(I_{\{d_k > 2a_{n_k}\}} \inf_u \mu_u^{(n_k)}(A_{n_k,k})). \quad \square
\end{aligned}$$

Proof of Lemma 3. (i) Let $\mu = \sup_i p(i) < 1$. By Markov's inequality,

$$\mathbf{P}\{\log_2[R_n p^n(x_1^n)] \geq \epsilon\sqrt{n} \,|\, x_1^n\} \leq \mathbf{E}\{R_n \,|\, x_1^n\} p^n(x_1^n) 2^{-\epsilon\sqrt{n}},$$

and by Kac's lemma (see, e.g., [26]) this is exactly equal to

$$[n + 1/p^n(x_1^n)]p^n(x_1^n)2^{-\epsilon\sqrt{n}} \leq [1 + n\mu^n]2^{-\epsilon\sqrt{n}} \leq C 2^{-\epsilon\sqrt{n}}$$

where $C = 1 + \sup_n\{n\mu^n\} < \infty$.

(ii) Write $K_n = 2^{-\epsilon\sqrt{n}}/p^n(x_1^n)$. By a simple union bound,

$$\begin{aligned}
\mathbf{P}\{\log_2[2np^n(x_1^n)] \leq \log_2[R_n p^n(x_1^n)] \leq -\epsilon\sqrt{n} \,|\, x_1^n\} &= \mathbf{P}\{2n \leq R_n \leq K_n \,|\, x_1^n\} \\
&\leq \sum_{j=2n}^{\lfloor K_n \rfloor} \mathbf{P}\{X_{j-n+1}^j = X_1^n \,|\, X_1^n = x_1^n\} \\
&\leq \lfloor K_n \rfloor p^n(x_1^n) \\
&\leq 2^{-\epsilon\sqrt{n}}.
\end{aligned}$$

(iii) By an application of the union bound as above, $\mathbf{P}\{n+1 \le R_n \le 2n-1\}$ is bounded above by

$$
\begin{aligned}
\sum_{j=1}^{n-1} \mathbf{P}\{X_{j+1}^{j+n} = X_1^n\} &= \sum_{j=1}^{\lfloor n/3 \rfloor} \mathbf{P}\{X_{j+1}^{j+n} = X_1^n\} + \sum_{j=\lfloor n/3 \rfloor+1}^{n-1} \mathbf{P}\{X_{j+1}^{j+n} = X_1^n\} \\
&= \sum_{j=1}^{\lfloor n/3 \rfloor} \sum_{x_1^j} p^j(x_1^j) \mathbf{P}\{X_{j+1}^{j+n} = X_1^n \mid X_1^j = x_1^j\} \\
&\quad + \sum_{j=\lfloor n/3 \rfloor+1}^{n-1} \sum_{x_{n-j+1}^n} p^j(x_{n-j+1}^n) \mathbf{P}\{X_{j+1}^{j+n} = X_1^n \mid X_{n-j+1}^n = x_{n-j+1}^n\} \\
&\le \sum_{j=1}^{\lfloor n/3 \rfloor} \mathbf{E}_{X_1^j} \left\{ [p^j(X_1^j)]^{\lfloor n/j \rfloor - 1} \right\} + \sum_{j=\lfloor n/3 \rfloor+1}^{n-1} \mathbf{E}_{X_{n-j+1}^n} \left\{ p^j(X_{n-j+1}^n) \right\}.
\end{aligned}
$$

With μ as in (i), we can bound both $p^j(X_1^j)$ and $p^j(X_{n-j+1}^n)$ by μ^j so that the above expression is at most

$$\sum_{j=1}^{\lfloor n/3 \rfloor} \mu^{n-2j} + \sum_{j=\lfloor n/3 \rfloor+1}^{n-1} \mu^j.$$

Summing these two geometric series we get an upper bound of the order of

$$(\text{const})\mu^{n/3} = \alpha 2^{-\beta n},$$

for appropriately chosen constants $\alpha, \beta > 0$. □

References

[1] A. Antos. *Performance limits of nonparametric estimators.* PhD thesis, Technical University of Budapest, H-1521 Stoczek u. 2, Budapest, Hungary, 1999.

[2] A. Antos. On nonparametric estimates of the expectation. In *Colloquium on Limit Theorems in Probability and Statistics, Abstracts of the Talks*, Balatonlelle, Hungary, 1999. BJMT.

[3] A. Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:643–645, 1999.

[4] B. Von Bahr and C.G. Esseen. Inequalities for the rth absolute moment of a sum of independent random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.*, 36:299–303, 1965.

[5] G.P. Bašarin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theor. Probability Appl.*, 4:333–336, 1959.

[6] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71:271–291, 1986.

[7] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[8] T. Cover. Rates of convergence for nearest neighbor procedures. *Proc. Hawaii Int. Conf. on Syst. Sci.*, Honolulu, HI, 413–415, 1968.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, New York, 1991.

[10] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:154–157, 1982.

[11] L. Devroye. On arbitrarily slow rates of global convergence in density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:475–483, 1983.

[12] L. Devroye. *A Course in Density Estimation.* Birkhäuser, Boston, 1987.

[13] L. Devroye. Exponential inequalities in nonparametric estimation. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, pages 31–44. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[14] L. Devroye. Another proof of a slow convergence result of Birgé. *Statistics and Probability Letters*, 23:63–67, 1995.

[15] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York, 1985.

[16] L. Devroye, L. Györfi and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[17] I. Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *J. Theoret. Probab.*, 11:795–811, 1998.

[18] I. Kontoyiannis, P.H. Algoet, Yu.M. Suhov, and A.J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory*, 44:1319–1327, 1998.

[19] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188, Cambridge University Press, Cambridge, 1989.

[20] V.V. Petrov. *Limit Theorems of Probability Theory.* Clarendon Press, Oxford, 1995.

[21] A. Rényi. *Probability Theory.* Akadémiai Kiadó, Budapest, 1970.

[22] T. Schürmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6:414–427, 1996.

[23] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[24] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences.* John Wiley, New York, 2001.

[25] S. Verdú. Fifty years of Shannon theory. *IEEE Trans. Inform. Theory*, 44: 2057–2078, 1998.

[26] A.D. Wyner, and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory*, 35:1250–1258, 1989.

[27] E.-H. Yang, and Y. Jia. Universal lossless coding of sources with large or unbounded alphabets. *Numbers, Information and Complexity*, (Ingo Althofer, et al, eds.), Kluwer Academic Publishers, pp. 421–442, 2000.

[28] J. Ziv, and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23:337–343, 1977.

[29] J. Ziv, and A. Lempel. Compression of individual sequences by variable rate coding. *IEEE Trans. Inform. Theory*, 24:530–536, 1978.