[9] L. M. Kaplan and C.-C. J. Kuo, "Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Haar basis," *IEEE Trans. Signal Processing*, vol. 41, pp. 3554–3562, Dec. 1993.
[10] R. W. Dijkerman and R. R. Mazumdar, "On the correlation structure of the wavelet coefficients of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1609–1612, Sept. 1994.
[11] A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, pp. 904–909, Mar. 1992.
[12] R. W. Dijkerman and R. R. Mazumdar, "Wavelet representations of stochastic processes and multiresolution stochastic models," *IEEE Trans. Signal Processing*, vol. 42, pp. 1640–1652, July 1994.
[13] W.-L. Hwang, "Estimation of fractional Brownian motion embedded in a noisy environment using nonorthogonal wavelets," *IEEE Trans. Signal Processing*, vol. 47, pp. 2211–2219, Aug. 1999.
[14] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[15] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.
[16] B. Jawerth and W. Sweldens, "An overview of wavelet based multiresolution analyzes," *SIAM Rev.*, vol. 36, pp. 377–412, 1994.
[17] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[18] J. Liu and P. Moulin, "Information-theoretic analysis of inter-scale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Processing*, vol. 10, pp. 1647–1658, Nov. 2001.

Estimating Entropy on m Bins Given Fewer Than m Samples

Liam Paninski

***Abstract*—Consider a sequence p_N of discrete probability measures, supported on m_N points, and assume that we observe N independent and identically distributed (i.i.d.) samples from each p_N. We demonstrate the existence of an estimator of the entropy, $H(p_N)$, which is consistent even if the ratio N/m_N is bounded (and, as a corollary, even if this ratio tends to zero, albeit at a sufficiently slow rate).**

***Index Terms*—Approximation theory, bias, consistency, distribution-free bounds, entropy, estimation.**

Earlier work has examined the problem of estimating the entropy of a discrete distribution p, with support on $m < \infty$ "bins," given N independent and identically distributed (i.i.d.) samples from p. It has long been recognized [1] that the crucial quantity in this estimation problem is the ratio N/m: if the number of samples is much greater than the number of bins, the estimation problem is easy, and *vice versa*. This correspondence concentrates on the hard part of this problem: how do we estimate the entropy when N/m_N is bounded? (To allow the precise statement of asymptotic results, it is convenient here to let $m = m_N$ depend on N; see [2] for motivation, a brief review, and some recent results.) We show that a consistent estimator of the entropy exists in this regime (thus proving the main conjecture of [2]); the most surprising implication of this result is that it is possible to accurately estimate the entropy on m bins, given N samples, even when N/m is small (provided that both N and m are sufficiently large). We give an existence proof of this result here; see [2] for a more constructive demonstration of an estimator which numerically appears to have this interesting and useful consistency property.

The entropy of a discrete distribution p is defined, as usual, as

$$H(p) = -\sum_{i=1}^{m} p_i \log p_i$$

where i indexes the support points of p, and the logarithm is taken to be natural. Our main result is as follows.

Theorem 1: Let $N/m_N \geq c > 0$, uniformly in N. Then there exists an estimator $\hat{H}_N$ for the entropy H which is uniformly consistent in mean square; that is,

$$E(\hat{H}_N - H)^2 < \epsilon(c, N)$$

with $\epsilon(c, N) \searrow 0$ as $N \to \infty$.

Note that the above statement is uniform over all distributions supported on m_N bins; the main practical implication, therefore, is that we can construct entropy estimators with surprisingly small "worst case" risk, given just m and N. We have as an easy corollary.

Corollary 2: There exists an estimator which is uniformly consistent even if $N/m_N \to 0$, sufficiently slowly.

More colloquially, we can estimate the entropy on m bins given fewer than m samples, as advertised. This is interesting in that it shows, in a sense, that the individual probabilities p need not be precisely estimated for the entropy estimate to be consistent.

On the other hand, in [2] we showed that N/m_N cannot decay faster than $N^{-\alpha}$, $\alpha > 0$, for consistency to hold, so the result is somewhat delicate. We present another partial converse here, indicating that not all functionals of the form $\sum_i f(p_i)$ can be estimated so easily, even for f smooth and vanishing at $p = 0$.

Proposition 3: Define the power sum

$$F(p) \equiv \sum_{i=1}^{m} p_i^{\alpha}, \qquad 0 < \alpha < 1.$$

If $\limsup N^{\alpha/(1-\alpha)}/m_N < \infty$, then

$$\liminf_N \inf_{\hat{F}_N} \max_p E(\hat{F}_N - F)^2 > 0$$

where the second infimum is taken over all possible estimators for F, and the maximum over all probability measures on m_N bins.

In particular, we need many more than m samples to estimate the power sum on m bins, whenever the exponent $\alpha \leq 1/2$. This result also quantifies the intuition that $F(p)$ becomes harder to estimate as α decreases (and, in fact, is impossible to estimate—in a "worst case" sense, at least—as $\alpha \to 0$, where we interpret $F(p)$ as counting the number of bins i for which $p_i > 0$).

The proof of the main theorem is built on ideas from [2]. Our estimator will be of the linear form

$$\hat{H}_{a,N} \equiv \sum_{j=0}^{N} a_{j,N} h_j$$

where the count statistics h_j are defined as

$$h_j \equiv \sum_{i=1}^{m} 1(n_i = j)$$

Manuscript received March 11, 2003; revised March 10, 2004. This work was supported by predoctoral and postdoctoral fellowships from the Howard Hughes Medical Institute.

The author is with the Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, U.K. (e-mail: liam@gatsby.ucl.ac.uk).

Communicated by A. B. Nobel, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2004.833360

with n_i the number of samples observed in bin i, and $a_N \equiv \{a_{j,N}\}_{0\le j\le N}$ is a set of $N+1$ scalars. To give a sense of what this linear form means, note that the obvious estimator for H, the maximum-likelihood estimator (MLE)

$$\hat{H}_{\mathrm{MLE}} = -\sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N}$$

can also be expressed as

$$\hat{H}_{\mathrm{MLE}} = \sum_{j=0}^{N} a_{MLE,j,N} h_j$$

with

$$a_{\mathrm{MLE},j,N} \equiv -\frac{j}{N} \log \frac{j}{N}.$$

More generally, $\hat{H}_{a,N} = \sum_{i=1}^{m} g_N(n_i)$, with $g_N(j) = a_{j,N}$.

For this class of estimators, we have some simple bounds on the bias and variance (derived in [3], [2]). We bound the variance $V(\hat{H}_{a,N})$ using McDiarmid's technique [4]

$$\max_p V(\hat{H}_{a,N}) < N \max_{0\le j<N} (a_{j+1,N} - a_{j,N})^2.$$

For the bias $B(\hat{H}_{a,N})$, we have the following approximation-theoretic bound:

$$\max_p |B(\hat{H}_{a,N})| \le m \max_{0\le x\le 1} \left| H(x) - \sum_j a_{j,N} B_{j,N}(x) \right|$$

where we have abbreviated the entropy function

$$H(x) = -x \log x$$

and the binomial functions

$$B_{j,N}(x) \equiv \binom{N}{j} x^j (1-x)^{N-j}.$$

Note that both of the above two bounds are distribution free, that is, uniform over all possible underlying m-ary distributions p; in addition, all maxima in sight are achieved, as can be shown by a straightforward compactness and continuity argument.

Now we only need to find some sequence a_N^* for which the above bound on the maximum bias $\max_p |B(\hat{H}_{a^*,N})|$ is $o(1)$ as $N \to \infty$, while the maximal difference $\max_{0\le j<N} |a_{j+1,N}^* - a_{j,N}^*|$, which controls the variance of $\hat{H}_{a^*,N}$, decreases sufficiently quickly. We start with a good guess at what a_N^* should be, then correct this original guess incrementally as N grows. Our starting point is

$$a'_{j,N} = a_{\mathrm{MLE},j,N} + \frac{1 - j/N}{2N};$$

this specific a'_N was derived in [2] and turns out to correspond to a version of the bias-corrected MLE introduced by [1]. The proof proceeds by showing that the bias function $H(x) - \sum_j a'_{j,N} B_{j,N}(x)$ converges, in a certain sense, to a manageable limit function g. Then we approximate this "leftover" function g with the binomial functions in such a way that the maximal difference remains under control.

Our argument is based on two lemmas; we state and prove the lemmas, then give the proof of the theorem.

Lemma 4 (Uniform Convergence of Rescaled Bias Function): Define the sequence of functions on the nonnegative real line

$$f_N(t) \equiv \begin{cases} N\left(H(t/N) - \sum_j a'_{j,N} B_{j,N}(t/N)\right), & 0 \le t \le N \\ 0, & t > N. \end{cases}$$

This sequence converges uniformly to a continuous function g that vanishes at infinity (that is, $g(t) \to 0$ as $t \to \infty$).

Proof: Writing out the definition of f_N on the interval $[0, N]$, we have

$$f_N = -t \log t - E_{N,t/N}(-j \log j) - \frac{1}{2} + \frac{t}{2N}$$

where $E_{N,y}(z)$ denotes the expectation of the real-valued function z on the integers, with respect to the binomial measure with parameters (N, y). The first and third terms are constant in N; the fourth is negligible for $t = o(N)$, and we will prove that the second converges uniformly to the mean of $-j \log j$ with respect to the Poisson measure with parameter t. Once this is established, a second-order expansion of $-t \log t$ is enough to demonstrate that the limit function obtained

$$g(t) \equiv -t \log t - \frac{1}{2} + \sum_{j\ge 0} \frac{e^{-t} t^j}{j!} j \log j$$

vanishes at infinity. (Continuity is readily apparent, since g is the uniform limit of a sequence of uniformly continuous functions.) Our strategy is to break the interval $t \in [0, \infty)$ into three parts: first, the compacta; second, the range $t \to \infty, t = o(N)$; finally, $t = (1-c)N, 0 \le c < 1$. (The definition of f_N takes care of the range $[N, \infty)$.)

To prove uniform convergence on compacta, first recall that each individual $B_{j,N}(t/N)$ converges uniformly on compacta to the corresponding "Poisson function" $e^{-t} t^j / j!$ (this is an obvious corollary of the uniform convergence on compacta of $(1 + t/N)^N$ to e^t). To show that f_N converges uniformly on compacta as well, it is enough to note that $|j \log j| \le j^2$ and $|B_{j,N}(t/N)| \le t^j / j!$; hence, the sum is bounded above by the sum of $j^2 t^j / j!$, and this last sum is clearly convergent in j, uniformly on compacta in t. This establishes uniform convergence of f_N on compacta, as can be seen by splitting our sum into a (convergent) finite part and a (uniformly small) infinite tail.

Next we need to rule out the possible existence of a sequence $t_N \to \infty$ along which convergence does not occur (that is, $f_N(t_N) \not\to g(t_N)$). We prove this by showing that $g(t) \to 0$ as $t \to \infty$ and $f_N(t_N) \to 0$ as $t_N \to \infty$, uniformly. We will derive both of these statements via the second-order expansion

$$j \log j = t \log t + (1 + \log t)(j - t) + \frac{1}{2t}(j - t)^2 + o(j - t)^2.$$

The proofs for $g(t)$ and $f_N(t_N)$ are quite similar; we begin with g. Take expectations of the first three terms of the above expansion and compare the mean and variance of a Poisson random variable with rate t. Now we need only show that the final term is negligible in expectation as $t \to \infty$. The sum defining this expectation is handled, as usual, in two parts (cf. [2, proof of Theorem 5]): a Taylor bound controls the contribution for $|j - t| < z(t)\sqrt{t}$ and tail probability inequalities control $|j - t| > z(t)\sqrt{t}$, where $z(t)$ is a function such that $z(t) \to \infty$ sufficiently slowly as $t \to \infty$. The Taylor bound follows by computing $H'''(t) \sim t^{-2}$, and leads to the condition

$$t^{3/2} z(t)^3 = o\left(t - z(t)\sqrt{t}\right)^2$$

ensuring that the contribution of the first component vanishes asymptotically. The tail bounds can be derived via the standard exponential tail inequalities [5] or more directly through a Stirling approximation; either method leads to the conclusion that, for t large, $e^{-t} t^j / j!$ decays like a Gaussian in j of mean and variance t, implying that algebraic growth, $z(t) \sim t^\alpha$, $\alpha > 0$, is sufficient to make the tail contribution

$$\sum_{|j-t|>z(t)\sqrt{t}} \frac{e^{-t} t^j}{j!} \left(j \log j - t \log t - (1 + \log t)(j - t) - \frac{1}{2t}(j - t)^2 \right)$$

$o(1)$ as $t \to \infty$. Choosing $\alpha \in (0, 1/6)$ completes the proof that g vanishes at infinity.

To prove that $f_N(t_N) \to 0$ for $t_N \to \infty$, $t_N = o(N)$, simply repeat the arguments of the preceding paragraph after replacing the

Poisson measure of rate t with the binomial-$(N, t_N/N)$ measure (and note that the final $t/2N$ term in the definition of f_N is negligible in this range of t_N); the mean and variance formulas, and the necessary exponential inequalities [5], are asymptotically equivalent.

To complete the proof of the lemma, we need only show that $f_N(t) \to 0$ uniformly for $t \in [(1-c)N, N]$, for any $0 \le c < 1$. The argument remains very similar (in particular, the last $o(j-t)^2$ term in the second-order expansion is handled the same way); the key difference is that now the final term in the definition of f_N is nonnegligible. Rather, this term is chosen to match the variance $cN(1-c)$ of the corresponding binomial measures. □

Lemma 5 (Density of "Poisson Polynomials"): Define $C_0(\Re^+)$ as the space of continuous real functions on the nonnegative real axis, vanishing at infinity, under the uniform metric. The linear span of the "Poisson functions" $\{e^{-t}t^j/j!\}_{j=0,1,\ldots}$ is dense in $C_0(\Re^+)$.

Proof: We apply the algebra version of the Stone–Weierstrass theorem [6]: let X be a compact Hausdorff space, $C(X;\Re)$ the algebra of continuous real functions on X under the sup-norm (recall that an algebra is a vector space closed under the usual pointwise multiplication of vectors), and A a subalgebra with the "two-point interpolation" property: for any $x, y \in X$, $a, b \in \Re$, and $\epsilon > 0$ there exists an $h \in A$ such that $|h(x)-a|, |h(y)-b| < \epsilon$. Then A is dense in $C(X;\Re)$.

First, we compactify $\Re^+$ in the obvious way, adding the point at infinity (this is acceptable by the vanishing-at-infinity condition on $C_0(\Re^+)$ and on the Poisson functions). Proving that the Poisson "polynomials"—finite linear combinations of the Poisson functions—interpolate points is easy, except for the point at infinity, but this is unnecessary by the definition of $C_0(\Re^+)$. The Poisson functions by themselves are not closed under multiplication, but they generate an algebra; this follows by uniformly approximating $e^{-\alpha t}t^k$ with Poisson polynomials (that is, finite sums of the Poisson functions), where k and α are integers, $k \ge 0$, $\alpha > 1$. This is possible by the density of the polynomials in the space of real continuous functions on the nonnegative real axis with finite weighted norm

$$\|f\|_{\Re^+,\infty,e^{-t}} \equiv \sup_{t\ge 0}\left(e^{-t}|f(t)|\right)$$

(see, e.g., [7, Ch. VI, especially p. 170]; note in particular that this density result is in terms of the weighted norm $\|.\|_{\Re^+,\infty,e^{-t}}$, not the original sup-norm). This follows since we can write

$$\left|e^{-\alpha t}t^k - \sum_j b_j e^{-t}t^j\right|$$

in the weighted polynomial approximation form

$$e^{-t}\left|e^{-\alpha_1 t}t^k - \sum_j b_j t^j\right|$$

with $\alpha_1 \equiv \alpha - 1 \ge 1$, and clearly $\|e^{-\alpha_1 t}t^k\|_{\Re^+,\infty,e^{-t}} < \infty$. □

Proof: (Theorem 1): The proof is a series of diagonalization arguments. We write out the basic idea first, then give the precise bounds below. To begin, it is clear that $\mathrm{Var}(\hat{H}_{a',N}) \to 0$ as $N \to \infty$. However, as emphasized in [2], when $\limsup N/m_N < \infty$, the maximal bias of $\hat{H}_{a',N}$ remains bounded away from zero. Nevertheless, Lemma 4 gives us good control over the bias function corresponding to $\hat{H}_{a',N}$: the sup-norm of this function is asymptotically equal to that of g, on a $1/N$ scale. In addition, the fact that g is continuous and vanishes at infinity means that we should be able to approximate the bias function well by just perturbing a few of the binomial terms $a'_{j,N}$, with each perturbation of small enough magnitude that the $o(1)$ asymptotic behavior of the variance is undisturbed. Lemma 5 ensures the success of this program, once we note that the binomial functions $B_{j,N}(t/N)$, in turn, asymptotically resemble the Poisson functions $e^{-t}t^j/j!$.

More precisely, we need to build a sequence of finite sums of binomial functions that converges uniformly to the bias limit function g. For any fixed j, the binomial function $B_{j,N}$ converges uniformly to the corresponding Poisson function. Therefore, if we wait for N to become large enough (this is the first diagonalization), it is enough to approximate g by Poisson polynomials. Choose P_k, a sequence of Poisson polynomials, converging uniformly to g; this is possible by the density lemma and the separability of $C_0(\Re^+)$. To each P_k corresponds a set of coefficients $b_k \equiv \{b_{j,k}\}_{j\ge 0}$ such that

$$P_k = \sum_{j\ge 0} b_{j,k}\frac{e^{-t}t^j}{j!};$$

by definition, b_k is an infinite sequence of reals, of which only a finite number of elements are nonzero for any fixed k. Now we can define $\{a^*_N\}_{N_k \le N < N_{k+1}}$ to be equal to $a'_N + \frac{1}{N}b_k$, where the addition of the vectors is defined in the obvious way, and the sequence $N_k \nearrow \infty$ will be specified below. The maximal bias of the resulting estimator $\hat{H}_{a^*,N}$ is $o(1)$, by construction.

Now, finally, to define the sequence N_k. First, trivially, N_k must be large enough to make the above vector addition sensible. More importantly, we need to choose the sequence N_k to increase quickly enough to satisfy our variance requirement

$$\max_{0\le j<N}(a^*_{j+1,N} - a^*_{j,N})^2 = o(1/N);$$

this will entail one simple last diagonalization argument.

Let us put all the pieces together. Define

$$\begin{aligned}
e^f_N &\equiv \sup_{t\ge 0}|f_N(t) - g(t)| \\
e_k &\equiv \sup_{t\ge 0}|g(t) - P_k(t)| \\
e_{j,N} &\equiv \sup_{t\ge 0}\left|B_{j,N}\left(\min(t/N,1)\right) - \frac{e^{-t}t^j}{j!}\right| \\
e'_N &\equiv N\max_{0\le j<N}(a'_{j+1,N} - a'_{j,N})^2
\end{aligned}$$

and

$$e^b_k \equiv \max_{j\ge 0}(b_{j+1,k} - b_{j,k})^2.$$

All but the last of these sequences tend to zero in N or k, while the last is finite for any fixed k. We have that

$$\max_p |B(\hat{H}_{a^*,N})| \le \frac{1}{c}\left(e^f_N + e_{k_N} + \sum_{j\le j(k_N)} e_{j,N}|b_{j,k_N}|\right)$$

where we define

$$j(k) \equiv \max\{j : b_{j,k} \ne 0\}$$

and

$$k_N \equiv \max\{k : N \ge N_k\}.$$

We have assumed here that $j(k_N) \le N$. Similarly

$$\max_p V(\hat{H}_{a^*,N}) \le 2(e'_N + N^{-1}e^b_{k_N})$$

since $(a+b)^2 \le 2(a^2+b^2)$ for any $a, b \in \Re$. Define $\epsilon'(c,N)$ by adding the second bound above to the square of the first. Clearly, $\epsilon'(c,N) \to 0$ if $k_N \to \infty$ slowly enough that

$$\limsup_N\left(N^{-1}e^b_{k_N} + \sum_{j\le j(k_N)} e_{j,N}|b_{j,k_N}|\right) = 0.$$

Finally, the $\epsilon(c,N)$ of the theorem can be defined by taking the least monotonically decreasing sequence that majorizes $\epsilon'(c,N)$. □

Proof: (Corollary 2): We need only to demonstrate the existence of some sequence $c_N \to 0$ for which we can guarantee the implication

$$N/m_N \ge c_N \Longrightarrow E(\hat{H}_N - H)^2 < \epsilon_N$$

for some sequence $\epsilon_N \to 0$. For example, given $\epsilon(c, N)$ from Theorem 1, we could define c_N inductively

$$c_1 = 1$$
$$c_{N+1} = \begin{cases} c_N, & \text{if } \epsilon(\frac{c_N}{2}, N) > \frac{c_N}{2} \\ \frac{c_N}{2}, & \text{otherwise.} \end{cases}$$

Clearly, $c_N \searrow 0$, since $\epsilon(c, N) \searrow 0$ for any fixed c. Defining ϵ_N as $\epsilon(1, N)$ for all N such that $c_N = 1$ and as c_N for all other N, the claim is proven. □

It is worth noting that the above proofs can be easily strengthened to almost sure convergence, given a suitably chosen probability space; see [2] for an example of such a probability space, along with the exponential tail probability inequalities (derived, again via McDiarmid) sufficient for the application of the Borel–Cantelli lemma.

Proof: (Proposition 3): We use what is perhaps the canonical approach from the minimax literature [8]: the idea is to find two points in parameter space, separated by some fixed distance $\epsilon > 0$, which are indistinguishable with some positive probability. More precisely, we will produce two sequences of probability measures $p_{0,N}$ and $p_{1,N}$, such that

$$\liminf_N |F(p_{0,N}) - F(p_{1,N})| > 0 \tag{1}$$

and

$$\limsup_N \|p_{0,N}^N - p_{1,N}^N\|_1 < 2 \tag{2}$$

where p^N denotes the product measure of p and $\|.\|_1$ the usual L^1 norm. That the existence of such a pair implies the stated claim is standard (see, e.g., [9]): consider the Bayesian problem of estimating F, given a prior placing mass $1/2$ on each of $p_{0,N}$ and $p_{1,N}$. Clearly, the best Bayesian estimator for this simple two-point problem has an average error bounded away from zero; the argument is completed by noting that this average error is necessarily less than the maximum error appearing in the statement of the proposition.

We let $p_{0,N}$ be the probability measure supported on bin 1, and $p_{1,N}$ be the simple perturbation

$$p_{1,N}(i) = \begin{cases} (1 - (m_N - 1)t_N), & i = 1 \\ t_N, & i > 1. \end{cases}$$

The sequence t_N is chosen to be the smallest positive sequence implying condition (1); clearly, the condition holds if

$$\liminf_N m_N^{1/\alpha} t_N > 0.$$

Now we examine the L^1 norm in condition (2). Since $p_{0,N}^N$ places all of its mass on the single point for which all observed data fall in bin 1, we have that this distance remains bounded away from 2 iff $p_{1,N}(1)^N$ remains bounded away from 0; in other words, condition (2) holds whenever

$$\limsup_N N m_N t_N < \infty.$$

Both of the above conditions can be met when N is allowed to grow no more quickly than $m_N^{(1-\alpha)/\alpha}$, as claimed. □

It remains unclear at the moment whether a result like Theorem 1 holds for the power sum $F(p)$, for values of $\alpha > 1/2$; neither the techniques of the above proof, nor those of Lemma 4, seem to generalize usefully to this case. For example, the obvious analog of Lemma 4 for the power sum does not hold: we have instead that

$$N^\alpha \left((t/N)^\alpha - \sum_j (j/N)^\alpha B_{j,N}(t/N) \right) \to g_\alpha(t)$$

for a continuous function $g_\alpha(t)$, and unfortunately this N^α convergence rate does not give us the control on a $1/N$ scale we need for the proof of Theorem 1; indeed, numerical experiments (data not shown) indicate that the computational approach employed in [2] for estimating the entropy does not lead to an estimator for the power sum that is consistent when N/m remains bounded. Closing this theoretical gap would be of some interest, as $F(p)$ approximates $H(p)$ in a sense for $\alpha \to 1$ [10].

ACKNOWLEDGMENT

We thank J. Victor for interesting conversations and the anonymous referees for many helpful comments.

REFERENCES

[1] G. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology II-B*. Glencoe, IL: Free Press, 1955, pp. 95–100.
[2] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1253, 2003.
[3] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures and Algorithms*, vol. 19, pp. 163–193, 2001.
[4] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[5] H. Chernoff, "A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–509, 1952.
[6] W. Rudin, *Functional Analysis*. New York: McGraw-Hill, 1973.
[7] P. Koosis, *The Logarithmic Integral*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[8] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.
[9] D. Donoho and R. Liu, "Geometrizing rates of convergence," *Ann. Statist.*, vol. 19, pp. 633–701, 1991.
[10] A. Renyi, "On measures of entropy and information," in *Proc. 4th Berkley Symp. Mathematical Statistics and Probability*, vol. 1, 1961, pp. 547–561.