ARTICLE Communicated by Jonathan Victor

Estimation of Entropy and Mutual Information

Liam Paninski
liam@cns.nyu.edu
Center for Neural Science, New York University, New York, NY 10003, U.S.A.

We present some new results on the nonparametric estimation of entropy and mutual information. First, we use an exact local expansion of the entropy function to prove almost sure consistency and central limit theorems for three of the most commonly used discretized information estimators. The setup is related to Grenander's method of sieves and places no assumptions on the underlying probability measure generating the data. Second, we prove a converse to these consistency theorems, demonstrating that a misapplication of the most common estimation techniques leads to an arbitrarily poor estimate of the true information, even given unlimited data. This "inconsistency" theorem leads to an analytical approximation of the bias, valid in surprisingly small sample regimes and more accurate than the usual $\frac{1}{N}$ formula of Miller and Madow over a large region of parameter space. The two most practical implications of these results are negative: (1) information estimates in a certain data regime are likely contaminated by bias, even if "bias-corrected" estimators are used, and (2) confidence intervals calculated by standard techniques drastically underestimate the error of the most common estimation methods.

Finally, we note a very useful connection between the bias of entropy estimators and a certain polynomial approximation problem. By casting bias calculation problems in this approximation theory framework, we obtain the best possible generalization of known asymptotic bias results. More interesting, this framework leads to an estimator with some nice properties: the estimator comes equipped with rigorous bounds on the maximum error over all possible underlying probability distributions, and this maximum error turns out to be surprisingly small. We demonstrate the application of this new estimator on both real and simulated data.

1 Introduction

The mathematical theory of information transmission represents a pinnacle of statistical research: the ideas are at once beautiful and applicable to a remarkably wide variety of questions. While psychologists and neurophysiologists began to apply these concepts almost immediately after their introduction, the past decade has seen a dramatic increase in the

popularity of information-theoretic analysis of neural data. It is unsurprising that these methods have found applications in neuroscience; after all, the theory shows that certain concepts, such as mutual information, are unavoidable when one asks the kind of questions neurophysiologists are interested in. For example, the capacity of an information channel is a fundamental quantity when one is interested in how much information can be carried by a probabilistic transmission system, such as a synapse. Likewise, we should calculate the mutual information between a spike train and an observable signal in the world when we are interested in asking how much we (or any homunculus) could learn about the signal from the spike train. In this article, we will be interested not in *why* we should estimate information-theoretic quantities (see Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Cover & Thomas, 1991 for extended and eloquent discussions of this question) but rather *how well* we can estimate these quantities at all, given finite independently and identically distributed (i.i.d.) data.

One would think this question would be well understood; after all, applied statisticians have been studying this problem since the first appearance of Shannon's papers, over 50 years ago (Weaver & Shannon, 1949). Somewhat surprisingly, though, many basic questions have remained unanswered. To understand why, consider the problem of estimating the mutual information, $I(X;Y)$, between two signals X and Y. This estimation problem lies at the heart of the majority of applications of information theory to data analysis; to make the relevance to neuroscience clear, let X be a spike train, or an intracellular voltage trace, and Y some behaviorally relevant, physically observable signal, or the activity of a different neuron. In these examples and in many other interesting cases, the information estimation problem is effectively infinite-dimensional. By the definition of mutual information, we require knowledge of the joint probability distribution $P(X,Y)$ on the range spaces of X and Y, and these spaces can be quite large—and it would seem to be very difficult to make progress here in general, given the limited amount of data one can expect to obtain from any physiological preparation.

In this article, we analyze a discretization procedure for reducing this very hard infinite-dimensional learning problem to a series of more tractable finite-dimensional problems. While we are interested particularly in applications to neuroscience, our results are valid in general for any information estimation problem. It turns out to be possible to obtain a fairly clear picture of exactly how well the most commonly used discretized information estimators perform, why they fail in certain regimes, and how this performance can be improved. Our main practical conclusions are that the most common estimators can fail badly in common data-analytic situations and that this failure is more dramatic than has perhaps been appreciated in the literature. The most common procedures for estimating confidence intervals, or error bars, fail even more dramatically in these "bad"

regimes. We suggest a new approach here and prove some of its advantages.

The article is organized as follows. In section 2, we define the basic regularization procedure (or, rather, formalize an intuitive scheme that has been widely used for decades). We review the known results in section 3 and then go on in section 4 to clarify and improve existing bias and variance results, proving consistency and asymptotic normality results for a few of the most commonly used information estimators. These results serve mainly to show when these common estimators can be expected to be accurate and when they should be expected to break down. Sections 5 and 6 contain the central results of this article. In section 5, we show, in a fairly intuitive way, why these common estimators perform so poorly in certain regimes and exactly how bad this failure is. These results lead us, in section 6, to study a polynomial approximation problem associated with the bias of a certain class of entropy estimators; this class includes the most common estimators in the literature, and the solution to this approximation problem provides a new estimator with much better properties. Section 7 describes some numerical results that demonstrate the relevance of our analysis for physiological data regimes. We conclude with a brief discussion of three extensions of this work: section 8.1 examines a surprising (and possibly useful) degeneracy of a Bayesian estimator, section 8.2 gives a consistency result for a potentially more powerful regularization method than the one examined in depth here, and section 8.3 attempts to place our results in the context of estimation of more general functionals of the probability distribution (that is, not just entropy and mutual information). We attach two appendixes. In appendix A, we list a few assorted results that are interesting in their own right but did not fit easily into the flow of the article. In appendix B, we give proofs of several of the more difficult results, deferred for clarity's sake from the main body of the text. Throughout, we assume little previous knowledge of information theory beyond an understanding of the definition and basic properties of entropy (Cover & Thomas, 1991); however, some knowledge of basic statistics is assumed (see, e.g., Schervish, 1995, for an introduction).

This article is intended for two audiences: applied scientists (especially neurophysiologists) interested in using information-theoretic techniques for data analysis and theorists interested in the more mathematical aspects of the information estimation problem. This split audience could make for a somewhat split presentation: the correct statement of the results requires some mathematical precision, while the demonstration of their utility requires some more verbose explanation. Nevertheless, we feel that the intersection between the applied and theoretical communities is large enough to justify a unified presentation of our results and our motivations. We hope readers will agree and forgive the length of the resulting article.

2 The Setup: Grenander's Method of Sieves

Much of the inherent difficulty of our estimation problem stems from the fact that the mutual information,

$$I(X, Y) \equiv \int_{\mathcal{X}\times\mathcal{Y}} dP(x, y) \log \frac{dP(x, y)}{d(P(x) \times P(y))},$$

is a nonlinear functional of an unknown joint probability measure, $P(X, Y)$, on two arbitrary measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. In many interesting cases, the "parameter space"—the space of probability measures under consideration—can be very large, even infinite-dimensional. For example, in the neuroscientific data analysis applications that inspired this work (Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998), $\mathcal{X}$ could be a space of time-varying visual stimuli and $\mathcal{Y}$ the space of spike trains that might be evoked by a given stimulus. This $\mathcal{Y}$ could be taken to be a (quite large) space of discrete (counting) measures on the line, while $\mathcal{X}$ could be modeled as the (even larger) space of generalized functions on $\Re^3$. Given N i.i.d. samples from $P(X, Y)$, $\{x_i, y_i\}_{1\le i\le N}$ ("stimulus" together with the evoked "response"), how well can we estimate the information this cell provides the brain about the visual scene? Clearly, it is difficult to answer this question as posed; the relationship between stimulus and response could be too complex to be revealed by the available data, even if N is large by neurophysiological standards. In fact, there are general theorems to this effect (section 3). Therefore, some kind of regularization is needed.

The most successful approach taken to date in our field to circumvent these problems was introduced by Bialek and colleagues (Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Strong et al., 1998). The idea is to admit to the difficulty of the problem and instead estimate a system of lower bounds on the mutual information via the data processing inequality (Cover & Thomas, 1991), which states that

$$I(X; Y) \ge I(S(X); T(Y)),$$

for any random variables X and Y and any functions S and T on the range of X and Y, respectively. The generality of the data processing inequality implies that we are completely unconstrained in our choice of S and T. So the strategy, roughly, is to choose a sequence of functions S_N and T_N that preserve as much information as possible given that $I(S_N; T_N)$ can be estimated with some fixed accuracy from N data samples. (Note that S_N and T_N are chosen independent of the data.) As the size of the available data set increases, our lower bound grows monotonically toward the true information. In slightly different language, S_N and T_N could be viewed as models, or parameterizations, of the allowed underlying measures $P(X, Y)$; we are simply allowing our model to become richer (higher-dimensional) as more data become available for fitting. Clearly, then, we are not intro-

ducing anything particularly novel, but merely formalizing what statisticians have been doing naturally since well before Shannon wrote his papers.

This strategy bears a striking resemblance to regularization methods employed in abstract statistical inference (Grenander, 1981), generally known as the *method of sieves*. Here, one replaces the parameter space of interest with a closely related space that simplifies the analysis or provides estimators with more attractive statistical properties. The following example is canonical and helps to clarify exactly why regularization is necessary. Say one is sampling from some unknown, smooth probability density function and is interested in estimating the underlying density. It is clear that there exists no maximum likelihood estimator of the density in the space of smooth functions (the object that formally maximizes the likelihood, a sum of Dirac point masses, does not lie in the allowed smoothness class). The situation is pathological, then: as the sample size increases to infinity, our estimate does not converge to the true density in the sense of any smooth topology. To avoid this pathology, we regularize our estimator by requiring that it take its values in a smooth function space. In effect, we restrict our attention to a subset, a "sieve," of the possible parameter space. As the available data increase, we gradually relax our constraints on the smoothness of the estimator (decrease the "mesh size" of our sieve), until in the limit our estimate of the underlying density is almost surely arbitrarily close to the true density. We will borrow this "mesh" and "sieve" terminology for the remainder of the article.

Here, we have to estimate a joint probability measure, $P(X, Y)$, on a large product space, $\mathcal{X} \times \mathcal{Y}$, in order to compute $I(X; Y)$. This is very difficult; therefore, we regularize our problem by instead trying to estimate $P(S, T)$ (where $P(S, T)$ is induced by the maps S and T in the natural way, i.e., $P(S = i, T = j) = P((x, y) : S(x) = i, T(y) = j)$). Thus, our "mesh size" is determined by the degree of compression inherent in going from (x, y) to $(S(x), T(y))$. Two variants of this strategy have appeared in the neuroscientific literature. The first, the so-called reconstruction technique (Bialek et al., 1991), makes use of some extremal property of the prior signal distribution to facilitate the reliable estimation of a lower bound on the true information. T_N here is a series of convolution operators, mapping spike trains (elements of $\mathcal{Y}$) back into the signal space $\mathcal{X}$. The lower bound on the information $I(X, T_N(Y))$ is estimated by spectral techniques: the prior distribution of X, $P(X)$, is chosen to be gaussian, and the well-known maximum-entropy property and spectral information formula for gaussian distributions provide the desired bound. The lower bounds obtained by this reconstruction approach have proven quite useful (Rieke et al., 1997); however, the available convergence results (of $I(X, T_N(Y))$ to $I(X, Y)$ as $N \to \infty$) rely on strong assumptions on $P(X, Y)$, and we will not discuss this technique in depth. (One final note: readers familiar with the reconstruction technique will realize that this example does not quite fit into our general framework, as the

convolution operators T_N, which are chosen by regression techniques, are in fact dependent on the data. These dependencies complicate the analysis significantly, and we will say very little on this topic beyond a brief note in section 8.2.)

The second method, the so-called direct method (Strong et al., 1998; Buracas, Zador, DeWeese, & Albright, 1998) is at first sight less dependent on assumptions on the prior distribtion on $\mathcal{X}$. Here one discretizes the space of all spike trains on some interval into some finite number, m, of words w, and makes use of the information formula for discrete distributions,

$$I(X; W) = H(W) - H(W \mid X),$$

to obtain a lower bound on the mutual information between the spike train and the signal of interest. $H(.)$ above denotes the entropy functional,

$$H(W) \equiv -\sum_i P(W_i) \log P(W_i),$$

and $H(. \mid .)$ denotes conditional entropy; X is, say, a visual signal on which we are conditioning.[1] In our previous notation, $W(y) = T(y)$. The generality of the data processing inequality, again, means that the discretization can take arbitrary form; letting T depend on the data size N, T_N could, for example, encode the total number of spikes emitted by the neuron for small N, then the occurrence of more detailed patterns of firing (Strong et al., 1998) for larger N, until, in the limit, all of the information in the spike train is retained.

Thus, in this "direct" approach, S_N and T_N are as simple as possible: these maps discretize $\mathcal{X}$ and $\mathcal{Y}$ into a finite number of points, $m_{S,N}$ and $m_{T,N}$, where $m_{S,N}$ and $m_{T,N}$ grow with N. For each value of N, our problem reduces to estimating $I(S_N, T_N)$, where the joint distribution of the random variables S_N and T_N is discrete on $m_{S,N}m_{T,N}$ points, and our parameter space, far from being infinite-dimensional, is the tractable $m_{S,N}m_{T,N}$-simplex, the set of convex combinations of $m_{S,N}m_{T,N}$ disjoint point masses. We emphasize again that neither S, T, nor m is allowed to depend on the data; in effect, we pretend that the discretizing maps and their ranges are chosen in advance, before we see a single sample.

While this discrete "binning" approach appears quite crude, it will allow us to state completely general strong convergence theorems for the information estimation problem, without any assumptions on, say, the ex-

[1] To keep data requirements manageable, $H(W \mid X)$—the expected conditional entropy of W given x, averaged over $P(X)$—is often replaced with $H(W \mid x)$, the conditional entropy given only a single x. The fact that any rigorous justification of this substitution requires a strong assumption (namely, that $H(W \mid x)$ is effectively independent of x with high $P(x)$-probability) has perhaps been overly glossed over in the literature.

istence or smoothness of a density for $P(X, Y)$. To our knowledge, results of this generality are unavailable outside the discrete context (but see Beirlant, Dudewicz, Gyorfi, & van der Meulen, 1997) for a good review of differential entropy estimation techniques, which provide a powerful alternative approach when the underlying probability measures are known a priori to possess a given degree of smoothness; Victor, 2002). In addition, of course, data that naturally take only a finite number of values are not uncommon. Therefore, we will analyze this discrete approach exclusively for the remainder of this article.

3 Previous Work

Most of the following results are stated in terms of the entropy $H(X)$; corresponding results for $I(X, Y)$ follow by Shannon's formula for discrete information:

$$I(X, Y) = H(X) + H(Y) - H(X, Y).$$

All of the estimators we will consider are functionals of the "empirical measures"

$$p_{N,i} \equiv \frac{1}{N} \sum_{j=1}^{N} \delta_i(T_N(y_j))$$

(where δ_i denotes the probability measure concentrated at i). The three most popular estimators for entropy seem to be:

1. The maximum likelihood (ML) estimator given p_N (also called the "plug-in"—by Antos & Kontoyiannis, 2001) or "naive"—by Strong et al., 1998—estimator),

   $$\hat{H}_{MLE}(p_N) \equiv -\sum_{i=1}^{m} p_{N,i} \log p_{N,i}$$

   (all logs are natural unless stated otherwise).

2. The MLE with the so-called Miller-Madow bias correction (Miller, 1955),

   $$\hat{H}_{MM}(p_N) \equiv \hat{H}_{MLE}(p_N) + \frac{\hat{m} - 1}{2N},$$

   where $\hat{m}$ is some estimate of the number of bins with nonzero P-probability (here we take $\hat{m}$ to be the number of bins with nonzero p_N-probability; see Panzeri & Treves, 1996, for some other examples).

3. The jackknifed (Efron & Stein, 1981) version of the MLE,

$$\hat{H}_{JK} \equiv N\hat{H}_{MLE} - \frac{N-1}{N}\sum_{j=1}^{N}\hat{H}_{MLE-j},$$

where H_{MLE-j} is the MLE based on all but the jth sample (unpublished notes of J. Victor; see also, e.g., Strong et al., 1998, in which a very similar estimator is used).

3.1 Central Limit Theorem, Asymptotic Bias and Variance. The majority of known results are stated in the following context: fix some discrete measure p on m bins and let N tend to infinity. In this case, the multinomial central limit theorem (CLT) implies that the empirical measures p_N are asymptotically normal, concentrated on an ellipse of size $\sim N^{-1/2}$ around the true discrete measure p; since $\hat{H}_{MLE}$ is a smooth function of p on the interior of the m-simplex, $\hat{H}_{MLE}$ is asymptotically normal (or chi-squared or degenerate, according to the usual conditions; Schervish, 1995) as well. It follows that both the bias and variance of $\hat{H}_{MLE}$ decrease approximately as $\frac{1}{N}$ (Basharin, 1959) at all but a finite number of points on the m-simplex. We will discuss this bias and variance rate explicitly for the above estimators in section 4; here it is sufficient to note that the asymptotic variance rate varies smoothly across the space of underlying probability measures $p(x, y)$, while the bias rate depends on only the number of nonzero elements of p (and is therefore constant on the interior of the m-simplex and discontinuous on the boundary). The asymptotic behavior of this estimation problem (again, when m is fixed and $N \to \infty$) is thus easily handled by classical techniques. While it does not seem to have been noted previously, it follows from the above that $\hat{H}_{MLE}$ is asymptotically minimax for fixed m as $N \to \infty$ (by "minimax," we mean best in a worst-case sense; we discuss this concept in more detail below); see Prakasa Rao (2001) for the standard technique, a clever "local Bayesian" application of the Cramer-Rao inequality.

Several articles (Miller, 1955; Carlton, 1969; Treves & Panzeri, 1995; Victor, 2000a) provide a series expansion for the bias, in the hope of estimating and subtracting out the bias directly. Although these authors have all arrived at basically the same answer, they have done so with varying degrees of rigor: for example, Miller (1955) uses an expansion of the logarithm that is not everywhere convergent (we outline this approach below and show how to avoid these convergence problems). Carlton (1969) rearranged the terms of a convergent expansion of the logarithm term in H; unfortunately, this expansion is not absolutely convergent, and therefore this rearrangement is not necessarily justified. Treves and Panzeri (1995) and Victor (2000a) both admit that their methods (a divergent expansion of the logarithm in each case) are not rigorous. Therefore, it would appear that none of the available results are strong enough to use in the context of this article, where

m and p can depend arbitrarily strongly on N. We will remedy this situation below.

3.2 Results of Antos and Kontoyiannis. Antos and Kontoyiannis (2001) recently contributed two relevant results. The first is somewhat negative:

Theorem (Antos & Kontoyiannis, 2001). *For any sequence* $\{\hat{H}_N\}$ *of entropy estimators, and for any sequence* $\{a_N\}, a_N \searrow 0$, *there is a distribution* P *on the integers* $\mathcal{Z}$ *with* $H \equiv H(P) < \infty$ *and*

$$\limsup_{n\to\infty} \frac{E(|\hat{H}_N - H|)}{a_N} = \infty.$$

In other words, there is no universal rate at which the error goes to zero, no matter what estimator we pick, even when our sample space is discrete (albeit infinite). Given any such putative rate a_N, we can always find some distribution P for which the true rate of convergence is infinitely slower than a_N. Antos and Kontoyiannis (2001) prove identical theorems for the mutual information, as well as a few other functionals of P.

The second result is an easy consequence of a more general fact about functions of multiple random variables; since we will use this general theorem repeatedly below, we reproduce the statement here. McDiarmid (1989) and Devroye, Gyorfi, and Lugosi (1996) provided a proof and extended discussions. The result basically says that if f is a function of N independent random variables, such that f depends only weakly on the value of any single variable, then f is tightly concentrated about its mean (i.e., $\mathrm{Var}(f)$ is small).

Theorem ("McDiarmid's inequality"; Chernoff, 1952; Azuma, 1967). *If* $\{x_j\}_{j:\,1,\ldots,N}$ *are independent random variables taking values in some arbitrary measurable space* A, *and* $f\colon A^N \mapsto \Re$ *is some function satisfying the coordinatewise boundedness condition,*

$$\sup_{\{x_1,\ldots,x_N\},x_j'} |f(x_1,\ldots,x_N) - f(x_1,\ldots,x_{j-1},x_j',x_{j+1},\ldots,x_N)| < c_j,$$
$$1 \le j \le N, \tag{3.1}$$

then, for any $\epsilon > 0$,

$$P(|f(x_1,\ldots,x_N) - E(f(x_1,\ldots,x_N))| > \epsilon) \le 2e^{-2\epsilon^2/\sum_{j=1}^N c_j^2}. \tag{3.2}$$

The condition says that by changing the value of the coordinate x_j, we cannot change the value of the function f by more than some constant c_j.

The usefulness of the theorem is a result of both the ubiquity of functions f satisfying condition 3.1 (and the ease with which we can usually check the condition) and the exponential nature of the inequality, which can be quite powerful if $\sum_{j=1}^{N} c_j^2$ satisfies reasonable growth conditions.

Antos and Kontoyiannis (2001) pointed out that this leads easily to a useful bound on the variance of the MLE for entropy:

Theorem (Antos & Kontoyiannis, 2001). *(a) For all N, the variance of the MLE for entropy is bounded above:*

$$Var(\hat{H}_{MLE}) \leq \left(\frac{(\log N)^2}{N} \right). \tag{3.3}$$

(b) Moreover, by McDiarmid's inequality, 3.2,

$$P(|\hat{H}_{MLE} - E(\hat{H}_{MLE})| > \epsilon) \leq 2e^{\frac{-N}{2}\epsilon^2 (\log N)^{-2}}. \tag{3.4}$$

Note that although this inequality is not particularly tight—while it says that the variance of $\hat{H}_{MLE}$ necessarily dives to zero with increasing N, the true variance turns out to be even smaller than the bound indicates—the inequality is completely universal, that is, independent of m or P. For example, Antos and Kontoyiannis (2001) use it in the context of m (countably) infinite. In addition, it is easy to apply this result to other functionals of p_N (see section 6 for one such important generalization).

3.3 $\hat{H}_{MLE}$ Is Negatively Biased Everywhere. Finally, for completeness, we mention the following well-known fact,

$$E_p(\hat{H}_{MLE}) \leq H(p), \tag{3.5}$$

where $E_p(.)$ denotes the conditional expectation given p. We have equality in the above expression only when $H(p) = 0$; in words, the bias of the MLE for entropy is negative everywhere unless the underlying distribution p is supported on a single point. This is all a simple consequence of Jensen's inequality; a proof was recently given in Antos and Kontoyiannis (2001), and we will supply another easy proof below. Note that equation 3.5 does not imply that the MLE for mutual information is biased upward everywhere, as has been claimed elsewhere; it is easy to find distributions p such that $E_p(\hat{I}_{MLE}) < I(p)$. We will discuss the reason for this misunderstanding below.

It will help to keep Figure 1 in mind. This figure gives a compelling illustration of perhaps the most basic fact about $\hat{H}_{MLE}$: the variance is small and the bias is large until $N \gg m$. This qualitative statement is not new; however, the corresponding quantitative statement—especially the fact that

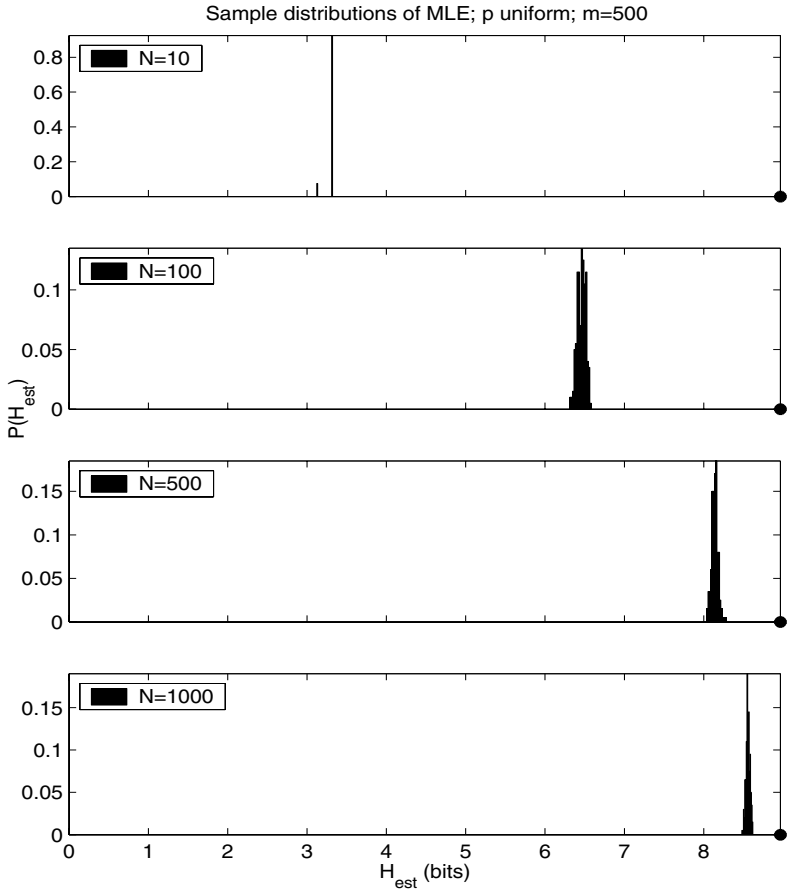

Figure 1: Evolution of sampling distributions of MLE: fixed m, increasing N. The true value of H is indicated by the dots at the bottom right corner of each panel. Note the small variance for all N and the slow decrease of the bias as $N \to \infty$.

$\hat{H}_{MLE}$ in the statement can be replaced with any of the three most commonly used estimators—appears to be novel. We will develop this argument over the next four sections and postpone discussion of the implications for data analysis until the conclusion.

4 The $N \gg m$ Range: The Local Expansion

The unifying theme of this section is a simple local expansion of the entropy functional around the true value of the discrete measure p, a variant of what

is termed the "delta method" in the statistics literature. This expansion is similar to one used by previous authors; we will be careful to note the extensions provided by the current work.

The main idea, outlined, for example, in Serfling (1980), is that any smooth function of the empirical measures p_N (e.g., any of the three estimators for entropy introduced above) will behave like an affine function with probability approaching one as N goes to infinity. To be more precise, given some functional f of the empirical measures, we can expand f around the underlying distribution p as follows:

$$f(p_N) = f(p) + df(p; p_N - p) + r_N(f, p, p_N),$$

where $df(p; p_N - p)$ denotes the functional derivative (Frechet derivative) of f with respect to p in the direction $p_N - p$, and $r_N(f, p, p_N)$ the remainder. If f is sufficiently smooth (in a suitable sense), the differential $df(p; p_N - p)$ will be a linear functional of $p_N - p$ for all p, implying

$$df(p; p_N - p) \equiv df\left(p; \frac{1}{N}\sum_{j=1}^{N} \delta_j - p\right) = \frac{1}{N}\sum_j df(p; \delta_j - p),$$

that is, $df(p; p_N - p)$ is the average of N i.i.d. variables, which implies, under classical conditions on the tail of the distribution of $df(p; \delta_j - p)$, that $N^{1/2} df(p; p_N - p)$ is asymptotically normal. If we can prove that $N^{1/2} r_N(f, p, p_N)$ goes to zero in probability (that is, the behavior of f is asymptotically the same as the behavior of a linear expansion of f about p), then a CLT for f follows. This provides us with a more flexible approach than the method outlined in section 3.1 (recall that that method relied on a CLT for the underlying empirical measures p_N, and such a CLT does not necessarily hold if m and p are not fixed).

Let us apply all this to H:

$$\begin{aligned}\hat{H}_{MLE}(p_N) &= H(p_N) \\ &= H(p) + dH(p; p_N - p) + r_N(H, p, p_N) \\ &= H(p) + \sum_{i=1}^{m} (p_i - p_{N,i}) \log p_i + r_N(H, p, p_N). \end{aligned} \tag{4.1}$$

A little algebra shows that

$$r_N(H, p, p_N) = -D_{KL}(p_N; p),$$

where $D_{KL}(p_N; p)$ denotes the Kullback-Leibler divergence between p_N, the empirical measure, and p, the true distribution. The sum in equation 4.1 has mean 0; by linearity of expectation, then,

$$E_p(\hat{H}_{MLE}) - H = -E_p(D_{KL}(p_N; p)), \tag{4.2}$$

and since $D_{KL}(p_N; p) \geq 0$, where the inequality is strict with positive probability whenever p is nondegenerate, we have a simple proof of the nonpositive bias of the MLE. Another (slightly more informative) approach will be given in section 6.

The second useful consequence of the local expansion follows by the next two well-known results (Gibbs & Su, 2002):

$$0 \leq D_{KL}(p_N; p) \leq \log(1 + \chi^2(p_N; p)), \tag{4.3}$$

where

$$\chi^2 \equiv \sum_{i=1}^{m} \frac{(p_{N,i} - p_i)^2}{p_i^2}$$

denotes Pearson's chi-square functional, and

$$E_p(\chi^2(p_N; p)) = \frac{|\text{supp}(p)| - 1}{N} \quad \forall p, \tag{4.4}$$

where $|\text{supp}(p)|$ denotes the size of the support of p, the number of points with nonzero p-probability. Expressions 4.2 through 4.4, with Jensen's inequality, give us rigorous upper and lower bounds on $B(\hat{H}_{MLE})$, the bias of the MLE:

Proposition 1.

$$-\log\left(1 + \frac{m-1}{N}\right) \leq B(\hat{H}_{MLE}) \leq 0,$$

with equality iff p is degenerate. The lower bound is tight as $N/m \to 0$, and the upper bound is tight as $N/m \to \infty$.

Here we note that Miller (1955) used a similar expansion to obtain the $\frac{1}{N}$ bias rate for m fixed, $N \to \infty$. The remaining step is to expand $D_{KL}(p_N; p)$:

$$D_{KL}(p_N; p) = \frac{1}{2}(\chi^2(p_N; p)) + O(N^{-2}), \tag{4.5}$$

if p is fixed. As noted in section 3.1, this expansion of D_{KL} does not converge for all possible values of p_N; however, when m and p are fixed, it is easy to show, using a simple cutoff argument, that this "bad" set of p_N has an asymptotically negligible effect on $E_p(D_{KL})$. The formula for the mean of the chi-square statistic, equation 4.4, completes Miller's and Madow's original proof (Miller, 1955); we have

$$B(\hat{H}_{MLE}) = -\frac{m-1}{2N} + o(N^{-1}), \tag{4.6}$$

if m is fixed and $N \to \infty$. From here, it easily follows that $\hat{H}_{MM}$ and $\hat{H}_{JK}$ both have $o(N^{-1})$ bias under these conditions (for $\hat{H}_{MM}$, we need only show that $\hat{m} \to m$ sufficiently rapidly, and this follows by any of a number of exponential inequalities [Dembo & Zeitouni, 1993; Devroye et al., 1996]; the statement for $\hat{H}_{JK}$ can be proven by direct computation). To extend these kinds of results to the case when m and p are not fixed, we have to generalize equation 4.6. This desired generalization of Miller's result does turn out to be true, as we prove (using a completely different technique) in section 6.

It is worth emphasizing that $E_p(\chi^2(p_N; p))$ is not constant in p; it is constant on the interior of the m-simplex but varies discontinuously on the boundary. This was the source of the confusion about the bias of the MLE for information,

$$\hat{I}_{MLE}(x, y) \equiv \hat{H}_{MLE}(x) + \hat{H}_{MLE}(y) - \hat{H}_{MLE}(x, y).$$

When $p(x, y)$ has support on the full $m_x m_y$ points, the $\frac{1}{N}$ bias rate is indeed given by $m_x m_y - m_x - m_y - 1$, which is positive for m_x, m_y large enough. However, $p(x, y)$ can be supported on as few as $\max(m_x, m_y)$ points, which means that the $\frac{1}{N}$ bias rate of $\hat{I}_{MLE}$ can be negative. It could be argued that this reduced-support case is nonphysiological; however, a simple continuity argument shows that even when $p(x, y)$ has full support but places most of its mass on a subset of its support, the bias can be negative even for large N, even though the asymptotic bias rate in this case is positive.

The simple bounds of proposition 1 form about half of the proof of the following two theorems, the main results of this section. They say that if $m_{S,N}$ and $m_{T,N}$ grow with N, but not too quickly, the "sieve" regularization works, in the sense that the sieve estimator is almost surely consistent and asymptotically normal and efficient on a $\sqrt{N}$ scale. The power of these results lies in their complete generality: we place no constraints whatsoever on either the underlying probability measure, $p(x, y)$, or the sample spaces $\mathcal{X}$ and $\mathcal{Y}$. Note that the theorems are true for all three of the estimators defined above (i.e., $\hat{H}$ above—and in the rest of the article, unless otherwise noted—can be replaced by $\hat{H}_{MLE}$, $\hat{H}_{JK}$, or $\hat{H}_{MM}$); thus, all three common estimators have the same $\frac{1}{N}$ variance rate: σ^2, as defined below. In the following, $\sigma_{X,Y}$ is the joint σ-algebra of $X \times Y$ on which the underlying probability distribution $p(X, Y)$ is defined, σ_{S_N,T_N} is the (finite) σ-algebra generated by S_N and T_N, and H_N denotes the N-discretized entropy, $H(S_N(X))$. The σ-algebra condition in theorem 1 is merely a technical way of saying that S_N and T_N asymptotically retain all of the data in the sample (x, y) in the appropriate measure-theoretic sense; see appendix A for details.

Theorem 1 (Consistency). *If $m_{S,N} m_{T,N} = o(N)$ and σ_{S_N,T_N} generates $\sigma_{X,Y}$, then $\hat{I} \to I$ a.s. as $N \to \infty$.*

Theorem 2 (Central limit). *Let*

$$\sigma_N^2 \equiv Var(-\log p_{T_N}) \equiv \sum_{i=1}^{m} p_{T_N,i}(-\log p_{T_N,i} - H_N)^2.$$

If $m_N \equiv m = o(N^{1/2})$, *and*

$$\liminf_{N\to\infty} N^{1-\alpha}\sigma_N^2 > 0$$

for some $\alpha > 0$, *then* $\left(\frac{N}{\sigma_N^2}\right)^{1/2}(\hat{H} - H_N)$ *is asymptotically standard normal.*

The following lemma is the key to the proof of theorem 1, and is interesting in its own right:

Lemma 1. *If* $m = o(N)$, *then* $\hat{H} \to H_N$ *a.s.*

Note that σ_N^2 in the statement of the CLT (theorem 2) is exactly the variance of the sum in expression 4.1, and corresponds to the asymptotic variance derived originally in Basharin (1959), by a similar local expansion. We also point out that σ_N^2 has a specific meaning in the theory of data compression (where σ_N^2 goes by the name of "minimal coding variance"; see Kontoyiannis, 1997, for more details).

We close this section with some useful results on the variance of $\hat{H}$. We have, under the stated conditions, that the variance of $\hat{H}$ is of order $\frac{\sigma^2}{N}$ asymptotically (by the CLT), and strictly less than $\frac{C\log(N)^2}{N}$ for all N, for some fixed C (by the result of Antos & Kontoyiannis 2001). It turns out that we can "interpolate," in a sense, between the (asymptotically loose but good for all N) p-independent bound and the (asymptotically exact but bad for small N) p-dependent gaussian approximation. The trick is to bound the average fluctuations in $\hat{H}$ when randomly replacing one sample, instead of the worst-case fluctuations, as in McDiarmid's bound. The key inequality is due to Steele (1986):

Theorem (Steele's inequality). *If* $S(x_1, x_2, \ldots, x_N)$ *is any function of* N *i.i.d. random variables, then*

$$var(S) \le \frac{1}{2}E\sum_{j=1}^{N}(S - S_i)^2,$$

where $S_i = S(x_1, x_2, \ldots, x_i', \ldots, x_N)$ *is given by replacing the* x_i *with an i.i.d. copy.*

For $S = \hat{H}$, it turns out to be possible to compute the right-hand side explicitly; the details are given in appendix B. It should be clear even without any computation that the bound so obtained is at least as good as the $\frac{C\log(N)^2}{N}$ guaranteed by McDiarmid; it is also easy to show, by the linear expansion technique employed above, that the bound is asymptotically tight under conditions similar to those of theorem 2.

Thus, $\sigma(p)^2$ plays the key role in determining the variance of $\hat{H}$. We know σ^2 can be zero for some p, since $\mathrm{Var}(-\log p_i)$ is zero for any p uniform on any k points, $k \leq m$. On the other hand, how large can σ^2 be? The following proposition provides the answer; the proof is in appendix B.

Proposition 2.

$$\max_p \sigma^2 \sim (\log m)^2.$$

This leads us to define the following bias-variance balance function, valid in the $N \gg m$ range:

$$V/B^2 \approx \frac{N(\log m)^2}{m^2}.$$

If V/B^2 is large, variance dominates the mean-square error (in the "worst-case" sense), and bias dominates if V/B^2 is small. It is not hard to see that if m is at all large, bias dominates until N is relatively huge (recall Figure 1). (This is just a rule of thumb, of course, not least because the level of accuracy desired, and the relative importance of bias and variance, depend on the application. We give more precise—in particular, valid for all values of N and m—formulas for the bias and variance in the following.)

To summarize, the sieve method is effective and the asymptotic behavior of $\hat{H}$ is well understood for $N \gg m$. In this regime, if $V/B^2 > 1$, classical (Cramer-Rao) effects dominate, and the three most common estimators ($\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$) are approximately equivalent, since they share the same asymptotic variance rate. But if $V/B^2 < 1$, bias plays a more important role, and estimators that are specifically designed to reduce the bias become competitive; previous work has demonstrated that $\hat{H}_{MM}$ and $\hat{H}_{JK}$ are effective in this regime (Panzeri & Treves, 1996; Strong et al., 1998). In the next section, we turn to a regime that is much more poorly understood, the (not uncommon) case when $N \sim m$. We will see that the local expansion becomes much less useful in this regime, and a different kind of analysis is required.

5 The $N \sim m$ Range: Consequences of Symmetry

The main result of this section is as follows: if N/m is bounded, the bias of $\hat{H}$ remains large while the variance is always small, even if $N \to \infty$. The

basic idea is that entropy is a symmetric function of $p_i, 1 \le i \le m$, in that H is invariant under permutations of the points $\{1, \ldots, m\}$. Most common estimators of H, including $\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$, share this permutation symmetry (in fact, one can show that there is some statistical justification for restricting our attention to this class of symmetric estimators; see appendix A). Thus, the distribution of $\hat{H}_{MLE}(p_N)$, say, is the same as that of $\hat{H}_{MLE}(p'_N)$, where p'_N is the rank-sorted empirical measure (for concreteness, define "rank-sorted" as "rank-sorted in decreasing order"). This leads us to study the limiting distribution of these sorted empirical measures (see Figure 2). It turns out that these sorted histograms converge to the "wrong" distribution under certain circumstances. We have the following result:

Theorem 3 (Convergence of sorted empirical measures; inconsistency). *Let P be absolutely continuous with respect to Lebesgue measure on the interval $[0, 1]$, and let $p = dP/dm$ be the corresponding density. Let S_N be the m-equipartition of $[0, 1]$, p' denote the sorted empirical measure, and $N/m \to c$, $0 < c < \infty$. Then:*

a. $p' \overset{L_1, a.s.}{\to} p'_{c,\infty}$, with $\|p'_{c,\infty} - p\|_1 > 0$. Here $p'_{c,\infty}$ is the monotonically decreasing step density with gaps between steps j and $j+1$ given by

$$\int_0^1 dt e^{-cp(t)} \frac{(cp(t))^j}{j!}.$$

b. Assume p is bounded. Then $\hat{H} - H_N \to B_{c,\hat{H}}(p)$ a.s., where $B_{c,\hat{H}}(p)$ is a deterministic function, nonconstant in p. For $\hat{H} = \hat{H}_{MLE}$,

$$B_{c,\hat{H}}(p) = h(p') - h(p) < 0,$$

where $h(.)$ denotes differential entropy.

In other words, when the sieve is too fine ($N \sim m$), the limit sorted empirical histogram exists (and is surprisingly easy to compute) but is not equal to the true density, even when the original density is monotonically decreasing and of step form. As a consequence, $\hat{H}$ remains biased even as $N \to \infty$. This in turn leads to a strictly positive lower bound on the asymptotic error of $\hat{H}$ over a large portion of the parameter space. The basic phenomenon is illustrated in Figure 2.

We can apply this theorem to obtain simple formulas for the asymptotic bias $B(p, c)$ for special cases of p: for example, for the uniform distribution $U \equiv U([0, 1])$,

$$B_{c,\hat{H}_{MLE}}(U) = \log(c) - e^{-c} \sum_{j=1}^{\infty} \frac{c^{j-1}}{(j-1)!} \log(j);$$

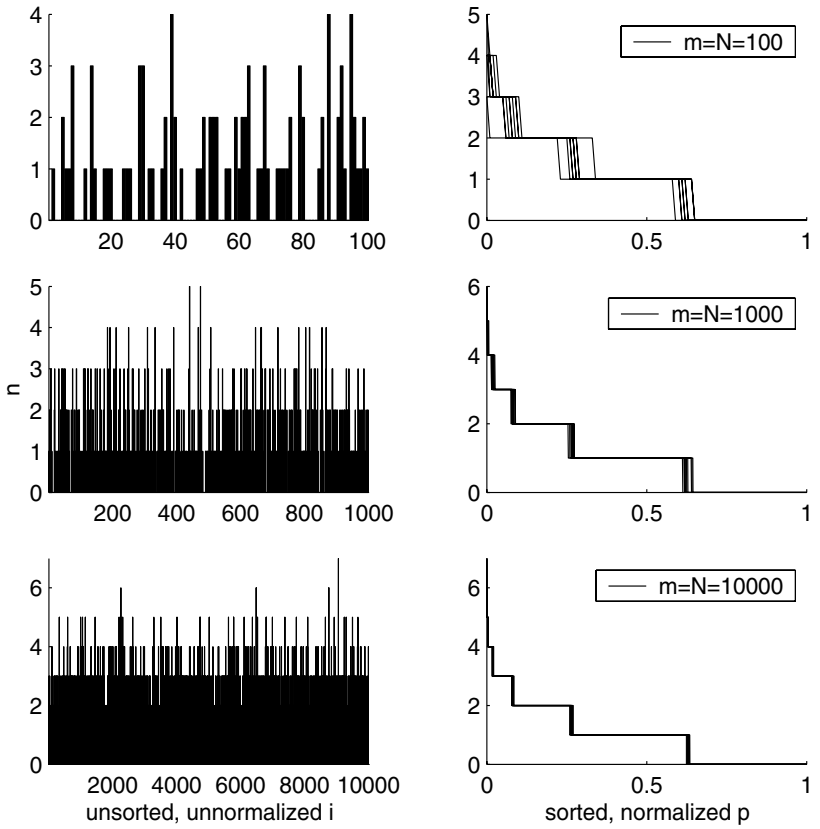

Figure 2: "Incorrect" convergence of sorted empirical measures. Each left panel shows an example unsorted m-bin histogram of N samples from the uniform density, with $N/m = 1$ and N increasing from top to bottom. Ten sorted sample histograms are overlaid in each right panel, demonstrating the convergence to a nonuniform limit. The analytically derived $p'_{c,\infty}$ is drawn in the final panel but is obscured by the sample histograms.

$$B_{c,\hat{H}_{MM}}(U) = B_{c,\hat{H}_{MLE}}(U) + \frac{1-e^{-c}}{2c};$$

$$B_{c,\hat{H}_{JK}}(U) = 1 + \log(c) - e^{-c}\sum_{j=1}^{\infty}\frac{c^{j-1}}{(j-1)!}(j-c)\log(j).$$

To give some insight into these formulas, note that $B_{c,\hat{H}_{MLE}}(U)$ behaves like $\log(N) - \log(m)$ as $c \to 0$, as expected given that $\hat{H}_{MLE}$ is supported on

$[0, \log(N)]$ (recall the lower bound of proposition 1); meanwhile,

$$B_{c,\hat{H}_{MM}}(U) \sim B_{c,\hat{H}_{MLE}}(U) + \frac{1}{2}$$

and

$$B_{c,\hat{H}_{JK}}(U) \sim B_{c,\hat{H}_{MLE}}(U) + 1$$

in this $c \to 0$ limit. In other words, in the extremely undersampled limit, the Miller correction reduces the bias by only half a nat, while the jackknife gives us only twice that. It turns out that the proof of the theorem leads to good upper bounds on the approximation error of these formulas, indicating that these asymptotic results will be useful even for small N. We examine the quality of these approximations for finite N and m in section 7.

This asymptotically deterministic behavior of the sorted histograms is perhaps surprising, given that there is no such corresponding deterministic behavior for the unsorted histograms (although, by the Glivenko-Cantelli theorem, van der Vaart & Wellner, 1996, there is well-known deterministic behavior for the integrals of the histograms). What is going on here? In crude terms, the sorting procedure "averages over" the variability in the unsorted histograms. In the case of the theorem, the "variability" at each bin turns out to be of a Poisson nature, in the limit as $m, N \to \infty$, and this leads to a well-defined and easy-to-compute limit for the sorted histograms.

To be more precise, note that the value of the sorted histogram at bin k is greater than t if and only if the number of (unsorted) $p_{N,i}$ with $p_{N,i} > t$ is at least k (remember that we are sorting in decreasing order). In other words,

$$p'_N = F_N^{-1},$$

where F_N is the empirical "histogram distribution function,"

$$F_N(t) \equiv \frac{1}{m} \sum_{i=1}^{m} 1(p_{N,i} < t),$$

and its inverse is defined in the usual way. We can expect these sums of indicators to converge to the sums of their expectations, which in this case are given by

$$E(F_N(t)) = \frac{1}{m} \sum_i P(p_{N,i} < t).$$

Finally, it is not hard to show that this last sum can be approximated by an integral of Poisson probabilities (see appendix B for details). Something similar happens even if $m = o(N)$; in this case, under similar conditions on p, we would expect each $p_{N,i}$ to be approximately gaussian instead of Poisson.

To compute $E(\hat{H})$ now, we need only note the following important fact: each $\hat{H}$ is a linear functional of the "histogram order statistics,"

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

where

$$n_i \equiv Np_{N,i}$$

is the unnormalized empirical measure. For example,

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

where

$$a_{\hat{H}_{MLE},j,N} = -\frac{j}{N} \log \frac{j}{N},$$

while

$$a_{\hat{H}_{JK},j,N} = N a_{\hat{H}_{MLE},j,N} - \frac{N-1}{N}((N-j)a_{\hat{H}_{MLE},j,N-1} + j a_{\hat{H}_{MLE},j-1,N-1}).$$

Linearity of expectation now makes things very easy for us:

$$\begin{aligned} E(\hat{H}) &= \sum_{j=0}^{N} a_{\hat{H},j,N} E(h_j) \\ &= \sum_j \sum_{i=1}^{m} a_{j,N} P(n_i = j) \\ &= \sum_j a_{j,N} \sum_i \binom{N}{j} p_i^j (1-p_i)^{N-j}. \end{aligned} \tag{5.1}$$

We emphasize that the above formula is exact for all N, m, and p; again, the usual Poisson or gaussian approximations to the last sum lead to useful asymptotic bias formulas. See appendix B for the rigorous computations.

For our final result of this section, let p and S_N be as in the statement of theorem 3, with p bounded, and $N = O(m^{1-\alpha})$, $\alpha > 0$. Then some easy computations show that $P(\exists i\colon\ n_i > j) \to 0$ for all $j > \alpha^{-1}$. In other words, with high probability, we have to estimate H given only $1 + \alpha^{-1}$ numbers, namely $\{h_j\}_{0 \le j \le \alpha^{-1}}$, and it is not hard to see, given equation 5.1 and the usual Bayesian lower bounds on minimax error rates (see, e.g., Ritov &

Bickel, 1990), that this is not enough to estimate $H(p)$. We have, therefore:

Theorem 4. *If $N \sim O(m^{1-\alpha})$, $\alpha > 0$, then no consistent estimator for H exists.*

By Shannon's discrete formula, a similar result holds for mutual information.

6 Approximation Theory and Bias

The last equality, expression 5.1, is key to the rest of our development. Letting $B(\hat{H})$ denote the bias of $\hat{H}$, we have:

$$\begin{aligned} B(\hat{H}) &= \left(\sum_{j=0}^{N} a_{j,N} \sum_{i=1}^{m} \binom{N}{j} p_i^j (1-p_i)^{N-j} \right) - \left(\sum_{i=1}^{m} -p_i \log(p_i) \right) \\ &= \left(\sum_i p_i \log(p_i) \right) + \sum_i \sum_j a_{j,N} \binom{N}{j} p_i^j (1-p_i)^{N-j} \\ &= \sum_i \left(p_i \log(p_i) + \sum_j a_{j,N} \binom{N}{j} p_i^j (1-p_i)^{N-j} \right). \end{aligned}$$

If we define the usual entropy function,

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

and the binomial polynomials,

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

we have

$$-B(\hat{H}) = \sum_i \left(H(p_i) - \sum_j a_{j,N} B_{j,N}(p_i) \right).$$

In other words, the bias is the m-fold sum of the difference between the function H and a polynomial of degree N; these differences are taken at the points p_i, which all fall on the interval $[0, 1]$. The bias will be small, therefore, if the polynomial is close, in some suitable sense, to H. This type of polynomial approximation problem has been extensively studied (Devore & Lorentz, 1993), and certain results from this general theory of approximation will prove quite useful.

Given any continuous function f on the interval, the Bernstein approximating polynomials of f, $B_N(f)$, are defined as a linear combination of the binomial polynomials defined above:

$$B_N(f)(x) \equiv \sum_{j=0}^{N} f(j/N) B_{j,N}(x).$$

Note that for the MLE,

$$a_{j,N} = H(j/N);$$

that is, the polynomial appearing in equation 5.1 is, for the MLE, exactly the Bernstein polynomial for the entropy function $H(x)$. Everything we know about the bias of the MLE (and more) can be derived from a few simple general facts about Bernstein polynomials. For example, we find the following result in Devore and Lorentz (1993):

Theorem (Devore & Lorentz, 1993, theorem 10.4.2). *If f is strictly concave on the interval, then*

$$B_N(f)(x) < B_{N+1}(f)(x) < f(x), \ 0 < x < 1.$$

Clearly, H is strictly concave, and $B_N(H)(x)$ and H are continuous, hence the bias is everywhere nonpositive; moreover, since

$$B_N(H)(0) = H(0) = 0 = H(1) = B_N(H)(1),$$

the bias is strictly negative unless p is degenerate. Of course, we already knew this, but the above result makes the following, less well-known proposition easy:

Proposition 3. *For fixed m and nondegenerate p, the bias of the MLE is strictly decreasing in magnitude as a function of N.*

(Couple the local expansion, equation 4.1, with Cover & Thomas, 1991, Chapter 2, Problem 34, for a purely information-theoretic proof.)

The second useful result is given in the same chapter:

Theorem (Devore & Lorentz, 1993, theorem 10.3.1). *If f is bounded on the interval, differentiable in some neighborhood of x, and has second derivative $f''(x)$ at x, then*

$$\lim_{N\to\infty} N(B_N(f)(x) - f(x)) = f''(x)\frac{x(1-x)}{2}.$$

This theorem hints at the desired generalization of Miller's original result on the asymptotic behavior of the bias of the MLE:

Theorem 5. *If $m > 1$, $N \min_i p_i \to \infty$, then*

$$\lim \frac{N}{m-1} B(\hat{H}_{MLE}) = -\frac{1}{2}.$$

The proof is an elaboration of the proof of the above theorem 10.3.1 of Devore and Lorentz (1993); we leave it for appendix B. Note that the convergence stated in the theorem given by Devore and Lorentz (1993) is not uniform for $f = H$, because $H(x)$ is not differentiable at $x = 0$; thus, when the condition of the theorem is not met (i.e., $\min_i p_i = O(\frac{1}{N})$), more intricate asymptotic bias formulas are necessary. As before, we can use the Poisson approximation for the bins with $Np_i \to c, 0 < c < \infty$ and an $o(1/N)$ approximation for those bins with $Np_i \to 0$.

6.1 "Best Upper Bounds" (BUB) Estimator. Theorem 5 suggests one simple way to reduce the bias of the MLE: make the substitution

$$\begin{aligned} a_{j,N} &= -\frac{j}{N} \log \frac{j}{N} \to a_{j,N} - H'' \left(\frac{j}{N} \right) \frac{\frac{j}{N}(1 - \frac{j}{N})}{2N} \\ &= -\frac{j}{N} \log \frac{j}{N} + \frac{(1 - \frac{j}{N})}{2N}. \end{aligned} \tag{6.1}$$

This leads exactly to a version of the Miller-Madow correction and gives another angle on why this correction fails in the $N \sim m$ regime: as discussed above, the singularity of $H(x)$ at 0 is the impediment.

A more systematic approach toward reducing the bias would be to choose $a_{j,N}$ such that the resulting polynomial is the best approximant of $H(x)$ within the space of N-degree polynomials. This space corresponds exactly to the class of estimators that are, like $\hat{H}$, linear in the histogram order statistics. We write this correspondence explicitly:

$$\{a_{j,N}\}_{0 \le j \le N} \longleftrightarrow \hat{H}_{a,N},$$

where we define $\hat{H}_{a,N} \equiv \hat{H}_a$ to be the estimator determined by $a_{j,N}$, according to

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

Clearly, only a small subset of estimators has this linearity property; the h_j-linear class comprises an $N+1$-dimensional subspace of the m^N-dimensional

space of all possible estimators. (Of course, m^N overstates the case quite a bit, as this number ignores various kinds of symmetries we would want to build into our estimator—see propositions 10 and 11—but it is still clear that the linear estimators do not exhaust the class of all reasonable estimators.) Nevertheless, this class will turn out to be quite useful.

What sense of "best approximation" is right for us? If we are interested in worst-case results, uniform approximation would seem to be a good choice: that is, we want to find the polynomial that minimizes

$$M(\hat{H}_a) \equiv \max_x \left| H(x) - \sum_j a_{j,N} B_{j,N}(x) \right|.$$

(Note the *the* above: the best approximant in this case turns out to be unique—Devore & Lorentz, 1993—although we will not need this fact below.) A bound on $M(\hat{H}_a)$ obviously leads to a bound on the maximum bias over all p:

$$\max_p |B(\hat{H}_a)| \leq mM(\hat{H}_a).$$

However, the above inequality is not particularly tight. We know, by Markov's inequality, that p cannot have too many components greater than $1/m$, and therefore the behavior of the approximant for x near $x = 1$ might be less important than the behavior near $x = 0$. Therefore, it makes sense to solve a weighted uniform approximation problem: minimize

$$M^*(f, \hat{H}_a) \equiv \sup_x \left(f(x) \left| H(x) - \sum_j a_{j,N} B_{j,N}(x) \right| \right),$$

where f is some positive function on the interval. The choice $f(x) = m$ thus corresponds to a bound of the form

$$\max_p |B(\hat{H}_a)| \leq c^*(f) M^*(f, \hat{H}_a),$$

with the constant $c^*(f)$ equal to one here. Can we generalize this?

According to the discussion above, we would like f to be larger near zero than near one, since p can have many small components but at most $1/x$ components greater than x. One obvious candidate for f, then, is $f(x) = 1/x$. It is easy to prove that

$$\max_p |B(\hat{H}_a)| \leq M^*(1/x, \hat{H}_a),$$

that is, $c^*(1/x) = 1$ (see appendix B). However, this f gives too much weight to small p_i; a better choice is

$$f(x) = \begin{cases} m & x < 1/m, \\ 1/x & x \geq 1/m. \end{cases}$$

For this f, we have:

Proposition 4.

$$\max_p |B(\hat{H}_a)| \leq c^*(f)M^*(f, \hat{H}_a)), \; c^*(f) = 2.$$

See appendix B for the proof.

It can be shown, using the above bounds combined with a much deeper result from approximation theory (Devore & Lorentz, 1993; Ditzian & Totik, 1987), that there exists an $a_{j,N}$ such that the maximum (over all p) bias is $O(\frac{m}{N^2})$. This is clearly better than the $O(\frac{m}{N})$ rate offered by the three most popular $\hat{H}$. We even have a fairly efficient algorithm to compute this estimator (a specialized descent algorithm developed by Remes; Watson, 1980). Unfortunately, the good approximation properties of this estimator are a result of a delicate balancing of large, oscillating coefficients $a_{j,N}$, and the variance of the corresponding estimator turns out to be very large. (This is predictable, in retrospect: we already know that no consistent estimator exists if $m \sim N^{1+\alpha}, \alpha > 0$.) Thus, to find a good estimator, we need to minimize bounds on bias and variance simultaneously; we would like to find $\hat{H}_a$ to minimize

$$\max_p (B_p(\hat{H}_a)^2 + V_p(\hat{H}_a)),$$

where the notation for bias and variance should be obvious enough. We have

$$\begin{aligned} \max_p (B_p(\hat{H}_a)^2 + V_p(\hat{H}_a)) &\leq \max_p B_p(\hat{H}_a)^2 + \max_p V_p(\hat{H}_a) \\ &\leq (c^*(f)M^*(f, \hat{H}_a))^2 + \max_p V_p(\hat{H}_a), \end{aligned} \tag{6.2}$$

and at least two candidates for easily computable uniform bounds on the variance. The first comes from McDiarmid:

Proposition 5.

$$Var(\hat{H}_a) < N \max_{0 \leq j < N} (a_{j+1} - a_j)^2.$$

This proposition is a trivial generalization of the corresponding result of Antos and Kontoyiannis (2001) for the MLE; the proofs are identical. We will make the abbreviation

$$\|Da\|_\infty^2 \equiv \max_{0 \leq j < N} (a_{j+1} - a_j)^2.$$

The second variance bound comes from Steele (1986); see appendix B for the proof (again, a generalization of the corresponding result for $\hat{H}$):

Proposition 6.

$$Var(\hat{H}_a) < 2c^*(f) \sup_x \left| f(x) \left(\sum_{j=2}^{N} j(a_{j-1} - a_j)^2 B_{j,N}(x) \right) \right|.$$

Thus, we have our choice of several rigorous upper bounds on the maximum expected error, over all possible underlying distributions p, of any given $\hat{H}_a$. If we can find a set of $\{a_{j,N}\}$ that makes any of these bounds small, we will have found a good estimator, in the worst-case sense; moreover, we will have uniform conservative confidence intervals with which to gauge the accuracy of our estimates. (Note that propositions 4 through 6 can be used to compute strictly conservative error bars for other h_j-linear estimators; all one has to do is plug in the corresponding $\{a_{j,N}\}$.)

Now, how do we find such a good $\{a_{j,N}\}$? For simplicity, we will base our development here on the McDiarmid bound, proposition 5, but very similar methods can be used to exploit the Steele bound. Our first step is to replace the above L_∞ norms with L_2 norms; recall that

$$M^*(f, H_a)^2 = \left\| (f) \left(H - \sum_j a_{j,N} B_{j,N} \right) \right\|_\infty^2.$$

So to choose $a_{N,j}$ in a computationally feasible amount of time, we minimize the following:

$$c^*(f)^2 \left\| (f) \left(H - \sum_j a_{j,N} B_{j,N} \right) \right\|_2^2 + N \|Da\|_2^2. \tag{6.3}$$

This is a "regularized least-squares" problem, whose closed-form solution is well known; the hope is that the (unique) minimizer of expression 6.3 is a near-minimizer of expression 6.2, as well. The solution for the best $a_{j,N}$, in vector notation, is

$$a = \left(X^t X + \frac{N}{c^*(f)^2} D^t D \right)^{-1} X^t Y, \tag{6.4}$$

where D is the difference operator, defined as in proposition 5, and $X^t X$ and $X^t Y$ denote the usual matrix and vector of self- and cross-products, $\langle B_{j,N} f, B_{k,N} f \rangle$ and $\langle B_{j,N} f, Hf \rangle$, respectively.

As is well known (Press, Teukolsky, Vetterling, & Flannery, 1992), the computation of the solution 6.4 requires on the order of N^3 time steps. We can improve this to an effectively $O(N)$-time algorithm with an empirical observation: for large enough j, the $a_{N,j}$ computed by the above algorithm look a lot like the $a_{N,j}$ described in expression 6.1 (data not shown). This is unsurprising, given Devore and Lorentz's theorem 10.3.1; the trick we took advantage of in expression 6.1 should work exactly for those j for which the function to be approximated is smooth at $x = \frac{j}{N}$, and $H(\frac{j}{N})$ becomes monotonically smoother as j increases.

Thus, finally, we arrive at an algorithm: for $0 < k < K \ll N$, set $a_{N,j} = -\frac{j}{N} \log \frac{j}{N} + \frac{(1-\frac{j}{N})}{2N}$ for all $j > k$, and choose $a_{N,j}, j \leq k$ to minimize the least-squares objective function 6.3; this entails a simple modification of equation 6.4,

$$a_{j \leq k} = \left(X^t X_{j \leq k} + \frac{N}{c^*(f)^2} (D^t D + I_k^t I_k) \right)^{-1} \left(X^t Y_{j \leq k} + \frac{N a_{k+1}}{c^*(f)^2} e_k \right),$$

where I_k is the matrix whose entries are all zero, except for a one at (k, k); e_k is the vector whose entries are all zero, except for a one in the kth element; $X^t X_{j \leq k}$ is the upper-left $k \times k$ submatrix of $X^t X$; and

$$X^t Y_{j \leq k} = \left\langle B_{j,N} f, \left(H - \sum_{j=k+1}^{N} a_{j,N} B_{j,N} \right) f \right\rangle.$$

Last, choose $a_{N,j}$ to minimize the true objective function, equation 6.2, over all K estimators so obtained. In practice, the minimal effective K varies quite slowly with N (for example, for $N = m < 10^5$, $K \approx 30$); thus, the algorithm is approximately (but not rigorously) $O(N)$. (Of course, once a good $\{a_{j,N}\}$ is chosen, $\hat{H}_a$ is no harder to compute than $\hat{H}$.) We will refer to the resulting estimator as $\hat{H}_{BUB}$, for "best upper bound" (Matlab code implementing this estimator is available on-line at http://www.cns.nyu.edu/∼liam).

Before we discuss $\hat{H}_{BUB}$ further, we note several minor but useful modifications of the above algorithm. First, for small enough N, the regularized least-squares solution can be used as the starting point for a hill-climbing procedure, minimizing expression 6.2 directly, for slightly improved results. Second, f, and the corresponding $c^*(f)^{-2}$ prefactor on the variance $(D^t D)$ term, can be modified if the experimenter is more interested in reducing bias than variance, or vice versa. Finally, along the same lines, we can constrain the size of a given coefficient $a_{k,N}$ by adding a Lagrange multiplier to the regularized least-square solution as follows:

$$\left(X^t X + \frac{N}{c^*(f)^2} D^t D + \lambda_k I_k^t I_k \right)^{-1} X^t Y,$$

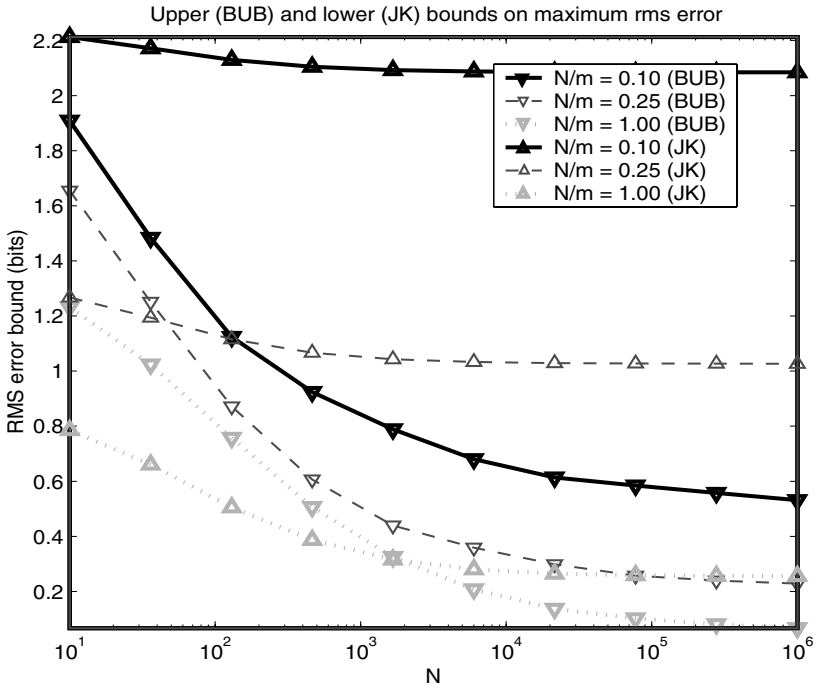

Figure 3: A comparison of lower bounds on worst-case error for $\hat{H}_{JK}$ (upward-facing triangles) to upper bounds on the same for $\hat{H}_{BUB}$ (downward-facing triangles), for several different values of N/m.

where I_k is as defined above. This is useful in the following context: at points p for which $H(p)$ is small, most of the elements of the typical empirical measure are zero; hence, the bias near these points is $\approx (N-1)a_{0,N} + a_{N,N}$, and λ_0 can be set as high as necessary to keep the bias as low as desired near these low entropy points. Numerical results show that these perturbations have little ill effect on the performance of the estimator; for example, the worst-case error is relatively insensitive to the value of λ_0 (see Figure 7).

The performance of this new estimator is quite promising. Figure 3 indicates that when m is allowed to grow linearly with N, the upper bound on the RMS error of this estimator (the square root of expression 6.2) drops off approximately as

$$\max_p((E(\hat{H}_{BUB} - H)^2)^{1/2}) <\sim N^{-\alpha}, \alpha \approx 1/3.$$

(Recall that we have a lower bound on the worst-case error of the three most common $\hat{H}$:

$$\max_p((E(\hat{H} - H)^2)^{1/2}) >\sim B_{\hat{H}}(N/m),$$

where $B_{\hat{H}}(N/m)$ is a bias term that remains bounded away from zero if N/m is bounded.) For emphasis, we codify this observation as a conjecture:

Conjecture. *$\hat{H}_{BUB}$ is consistent as $N \to \infty$ even if $N/m \sim c, 0 < c < \infty$.*

This conjecture is perhaps not as surprising as it appears at first glance; while, intuitively, the nonparametric estimation of the full distribution p on m bins should require $N \gg m$ samples, it is not a priori clear that estimating a single parameter, or functional of the distribution, should be so difficult. Unfortunately, while we have been able to sketch a proof of the above conjecture, we have not yet obtained any kind of complete asymptotic theory for this new estimator along the lines of the consistency results of section 4; we hope to return to this question in more depth in the future (see section 8.3).

From a nonasymptotic point of view, the new estimator is clearly superior to the three most common $\hat{H}$, even for small N, if N/m is small enough: the upper bounds on the error of the new estimator are smaller than the lower bounds on the worst-case error of $\hat{H}_{JK}$ for $N/m = 1$, for example, by $N \approx 1000$, while the crossover point occurs at $N \approx 50$ for $m = 4N$. (We obtain these lower bounds by computing the error on a certain subset of the parameter space on which exact calculations are possible; see section 7. For this range of N and m, $\hat{H}_{JK}$ always had a smaller maximum error than $\hat{H}_{MLE}$ or $\hat{H}_{MM}$.) For larger values of N/m or smaller values of N, the figure is inconclusive, as the upper bounds for the new estimator are greater than the lower bounds for $\hat{H}_{JK}$. However, the numerical results in the next section indicate that, in fact, $\hat{H}_{BUB}$ performs as well as the three most common $\hat{H}$ even in the $N \gg m$ regime.

7 Numerical Results and Applications to Data

What is the best way to quantify the performance of this new estimator (and to compare this performance to that of the three most common $\hat{H}$)? Ideally, we would like to examine the expected error of a given estimator simultaneously for all parameter values. Of course, this is possible only when the parameter space is small enough; here, our parameter space is the $(m - 1)$-dimensional space of discrete distributions on m points, so we can directly display the error function only if $m \leq 3$ (see Figure 4). For larger m, we can either compute upper bounds on the worst-case error, as in the previous section (this worst-case error is often considered the most

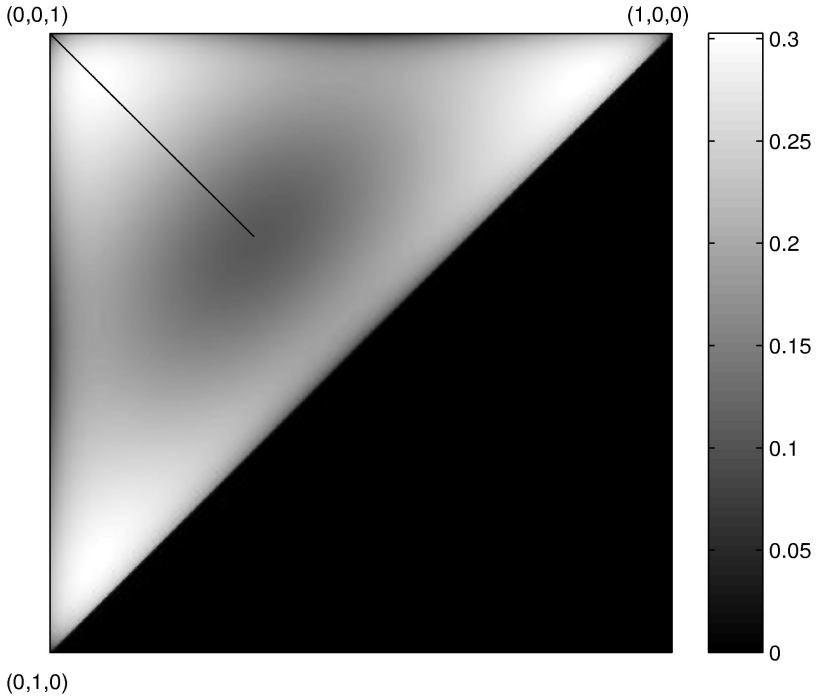

Figure 4: Exact RMS error surface (in bits) of the MLE on the 3-simplex, $N = 20$. Note the six permutation symmetries. One of the "central lines" is drawn in black.

important measure of an estimator's performance if we know nothing about the a priori likelihood of the underlying parameter values), or we can look at the error function on what we hope is a representative slice of the parameter space.

One such slice through parameter space is given by the "central lines" of the m-simplex: these are the subsets formed by linearly interpolating between the trivial (minimal entropy) and flat (maximal entropy) distributions (there are m of these lines, by symmetry). Figure 4 shows, for example, that the worst-case error for the MLE is achieved on these lines, and it seems plausible that these lines might form a rich enough class that it is as difficult to estimate entropy on this subset of the simplex as it is on the entire parameter space. While this intuition is not quite correct (it is easy to find reasonable estimators whose maximum error does not fall on these lines), calculating the error on these central lines does at least give us a lower bound on the worst-case error. By recursively exploiting the permutation symmetry and the one-dimensional nature of the problem, we constructed a fast

algorithm to compute these central line error functions exactly—explicitly enumerating all possible sorted histograms for a given (m, N) pair via a special recursion, computing the multinomial probability and estimating the $\hat{H}(p')$ associated with each histogram, and obtaining the desired moments of the error distribution at each point along the central line. The results are shown in Figures 5 and 6.

Figure 5 illustrates two important points. First, the new estimator performs quite well; its maximum error on this set of distributions is about half as large as that of the next best estimator, $\hat{H}_{JK}$, and about a fifth the size of the worst-case error for the MLE. In addition, even in the small region where the error of $\hat{H}$ is less than that of $\hat{H}_{BUB}$—near the point at which $H = \hat{H} = 0$—the error of the new estimator remains acceptably small. Second, these exact computations confirm the validity of the bias approximation of theorem 3, even for small values of N. Compare, for example, the bias predicted by the fixed m, large N theory (Miller, 1955), which is constant on the interior of this interval. This figure thus clearly shows that the classical asymptotics break down when the $N \gg m$ condition is not satisfied and that the $N \sim m$ asymptotics introduced in section 5 can offer a powerful replacement. Of course, neither approximation is strictly "better" than the other, but one could argue that the $N \sim m$ situation is in fact the more relevant for neuroscientific applications, where m is often allowed to vary with N.

In Figure 6, we show these central line error curves for a few additional (N, m) combinations. Recall Figure 3: if N is too small and N/m is too large, the upper bound on the error of $\hat{H}_{BUB}$ is in fact greater than the lower bound on the worst-case error for $\hat{H}_{JK}$; thus, the analysis presented in the previous section is inconclusive in this (N, m) regime. However, as Figure 6 indicates, the new estimator seems to perform well even as N/m becomes large; the maximum error of $\hat{H}_{BUB}$ on the central lines is strictly less than that of the three most common estimators for all observed combinations of N and m, even for $N = 10m$. Remember that all four estimators are basically equivalent as $n_i \to \infty$, where the classical (Cramer-Rao) behavior takes over and variance dominates the mean-square error of the MLE. In short, the performance of the new estimator seems to be even better than the worst-case analysis of section 6.1 indicated.

While the central lines are geometrically appealing, they are certainly not the only family of distributions we might like to consider. We examine two more such families in Figure 7 and find similar behavior. The first panel shows the bias of the same four estimators along the flat distributions on m' bins, $1 \leq m' \leq m$, where, as usual, only m and N are known to the estimator. Note the emergence of the expected log-linear behavior of the bias of $\hat{H}$ as N/m becomes small (recall the discussion following theorem 3). The second panel shows the bias along the family $p_i \simeq i^{\alpha}$, for $0 < \alpha < 20$, where similar behavior is evident. This figure also illustrates the effect of varying the λ_0 parameter: the bias at low entropy points can be reduced

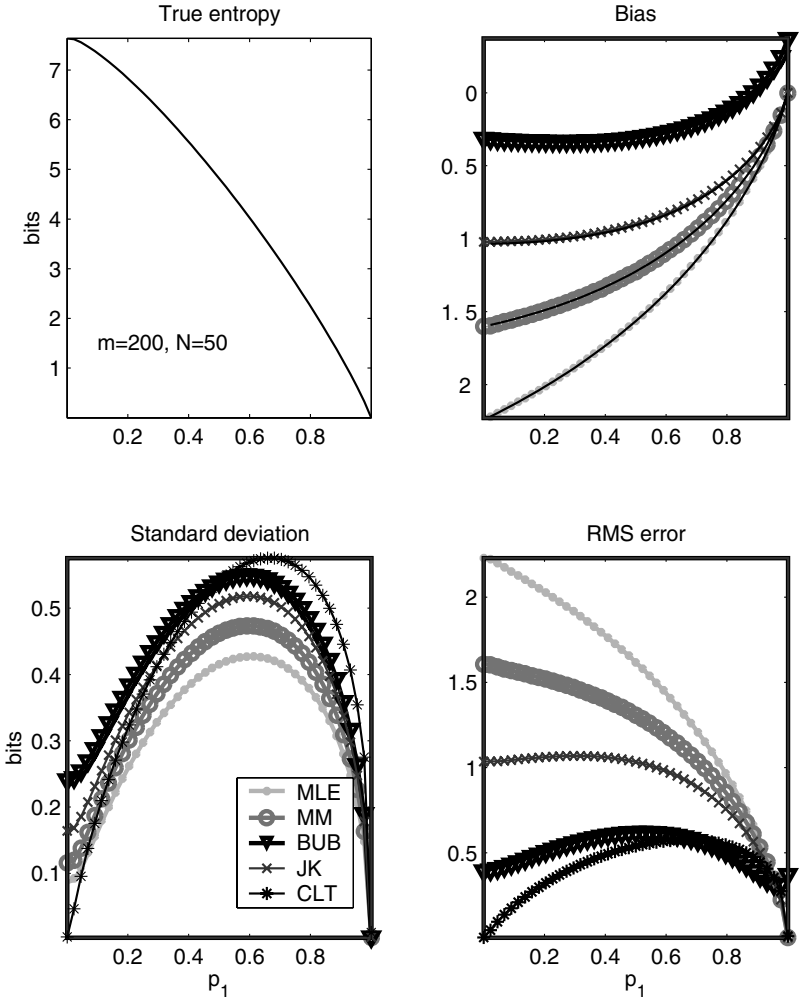

Figure 5: Example of error curves on the "central lines" for four different estimators ($N = 50$, $m = 200$; $\lambda_0 = 0$ here and below unless stated otherwise). The top left panel shows the true entropy, as p_1 ranges from 1 (i.e., p is the unit mass on one point) to $\frac{1}{m}$ (where p is the flat measure on m points). Recall that on the central lines, $p_i = \frac{1-p_1}{m-1} \forall i \neq 1$. The solid black lines overlying the symbols in the bias panel are the biases predicted by theorem 3. These predictions depend on N and m only through their ratio, N/m. The black dash-asterisk denotes the variance predicted by the CLT, $\sigma(p)N^{-1/2}$.

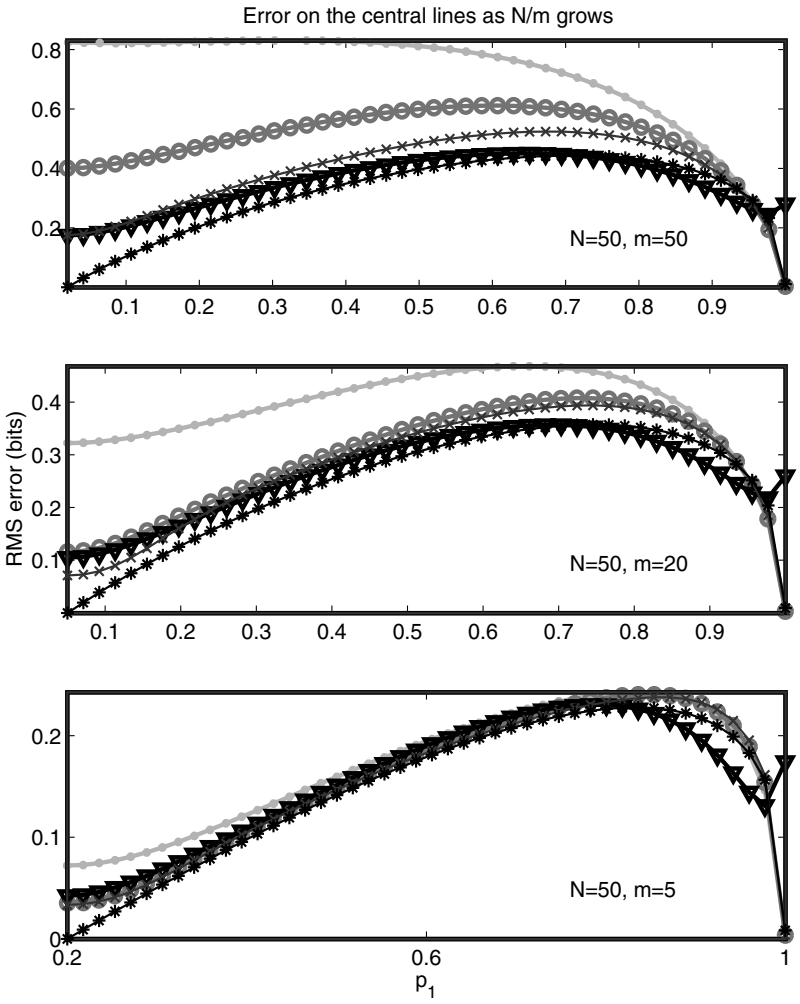

Figure 6: "Central line" error curves for three different values of N/m (notation as in Figure 5). Note that the worst-case error of the new estimator is less than that of the three most common $\hat{H}$ for all observed (N, m) pairs and that the error curves for the four estimators converge to the CLT curve as $N/m \to \infty$.

to arbitrarily low levels at the cost of relatively small changes in the bias at the high-entropy points on the m-simplex. As above, the Steele bounds on the variance of each of these estimators were comparable, with $\hat{H}_{BUB}$ making a modest sacrifice in variance to achieve the smaller bias shown here.

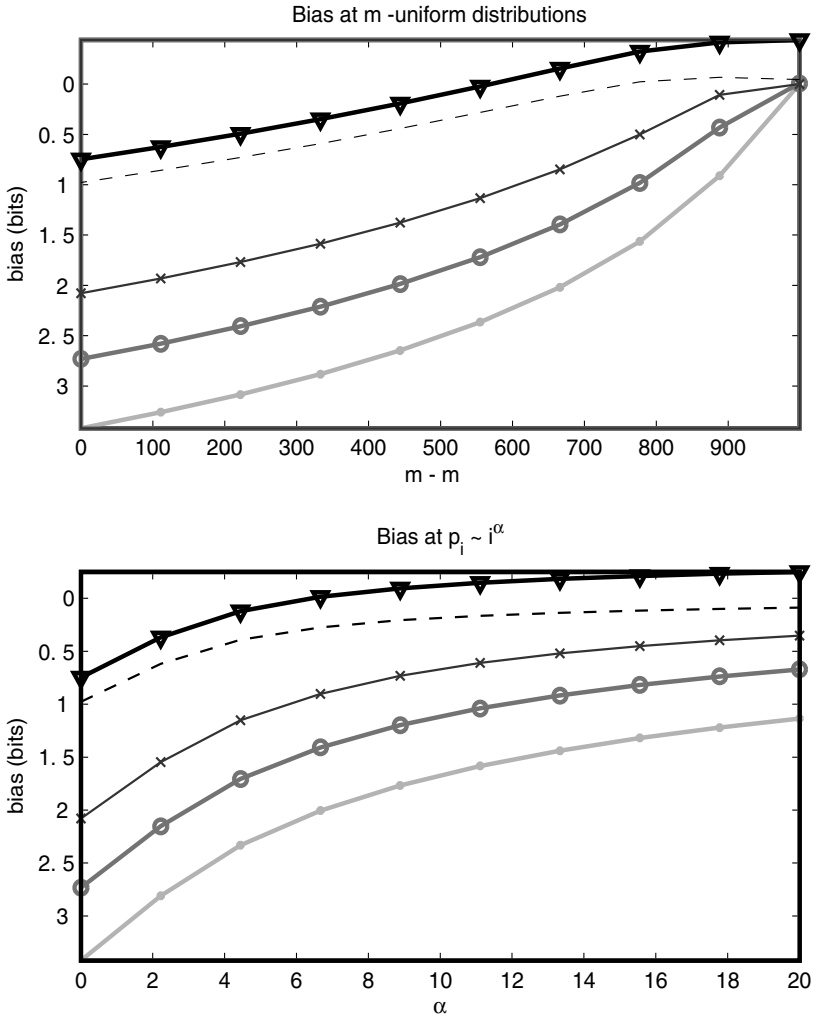

Figure 7: Exact bias on two additional families of distributions. The notation is as in Figure 5; the dashed line corresponds to $\hat{H}_{BUB}$ with λ_0 set to reduce the bias at the zero-entropy point. (Top) Flat distributions on m' bins, $1 \leq m' \leq m$. (Bottom) $p_i \simeq i^{\alpha}$. $N = 100$ and $m = 1000$ in each panel.

One could object that the set of probability measures examined in Figures 5, 6, and 7 might not be relevant for neural data; it is possible, for example, that probability measures corresponding to cellular activity lie in a completely different part of parameter space. In Figure 8, therefore, we examined our estimators' behavior over a range of p generated by the most commonly used neural model, the integrate-and-fire (IF) cell. The exact calculations presented in the previous figures are not available in this context, so we turned to a Monte Carlo approach. We drove an IF cell with i.i.d. samples of gaussian white noise, discretized the resulting spike trains in binary fashion (with discretization parameters comparable to those found in the literature), and applied the four estimators to the resulting binned spike trains.

Figure 8 shows the bias, variance, and root mean square error of our four estimators over a range of parameter settings, in a spirit similar to that of Figure 5; the critical parameter here was the mean firing rate, which was adjusted by systematically varying the DC value of the current driving the cell. (Because we are using simulated data, we can obtain the "true" value of the entropy simply by increasing N until $\hat{H}$ is guaranteed to be as close as desired to the true H, with probability approaching one.) Note that as the DC current increases, the temporal properties of the spike trains change as well; at low DC, the cells are essentially noise driven and have a correspondingly randomized spike train (as measured, e.g., by the coefficient of variation of the interspike interval distribution), while at high DC, the cells fire essentially periodically (low interspike interval coefficient of variation). The results here are similar to those in the previous two figures: the bias of the new estimator is drastically smaller than that of the other three estimators over a large region of parameter space. Again, when $H(p) \to 0$ (this occurs in the limit of high firing rates—when all bins contain at least one spike—and low firing rates, where all bins are empty), the common estimators outperform $\hat{H}_{BUB}$, but even here, the new estimator has acceptably small error.

Finally, we applied our estimators to two sets of real data (see Figures 9 and 10). The in vitro data set in Figure 9 was recorded in the lab of Alex Reyes. In a rat cortical slice preparation, we obtained double whole-cell patches from single cells. We injected a white-noise current stimulus via one electrode while recording the voltage response through the other electrode. Recording and data processing followed standard procedures (see Paninski, Lau, & Reyes, in press, for more detail). The resulting spike trains were binned according to the parameters given in the figure legend, which were chosen, roughly, to match values that have appeared in the literature. Results shown are from multiple experiments on a single cell; the standard deviation of the current noise was varied from experiment to experiment to explore different input ranges and firing rates. The in vivo data set in Figure 10 was recorded in the lab of John Donoghue. We recorded simultaneously from multiple cells in the arm representation of the primary motor cortex

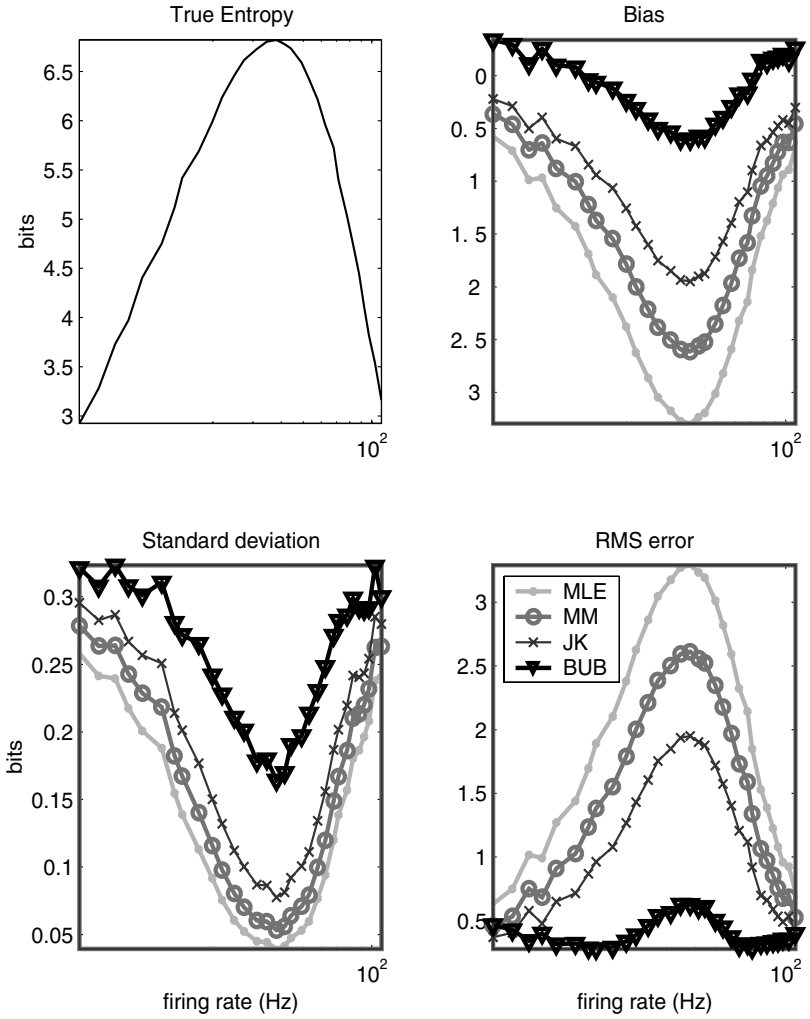

Figure 8: Error curves for simulated data (IF model, driven by white noise current), computed via Monte Carlo. $N = 100$ i.i.d. spike trains, time window $T = 200$ ms, binary discretization, bin width $dt = 20$ ms, thus, $m = 2^{10}$; DC of input current varied to explore different firing rates. Note the small variance and large negative bias of $\hat{H}$ over a large region of parameter space. The variance of $\hat{H}_{BUB}$ is slightly larger, but this difference is trivial compared to the observed differences in bias.

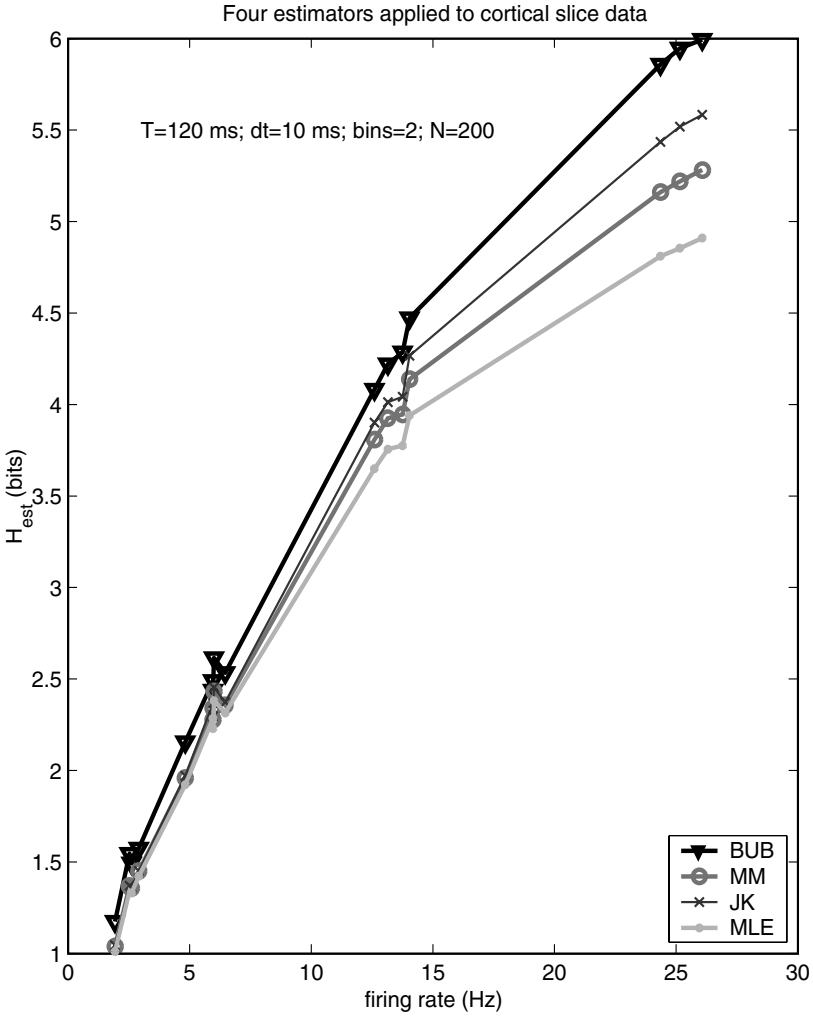

Figure 9: Estimated entropy of spike trains from a single cell recorded in vitro. The cell was driven with a white noise current. Each point corresponds to a single experiment, with $N = 200$ i.i.d. trials. The standard deviation of the input noise was varied from experiment to experiment. Spike trains were 120 ms long, discretized into 10 ms bins of 0 or 1 spike each; $m = 2^{12}$.

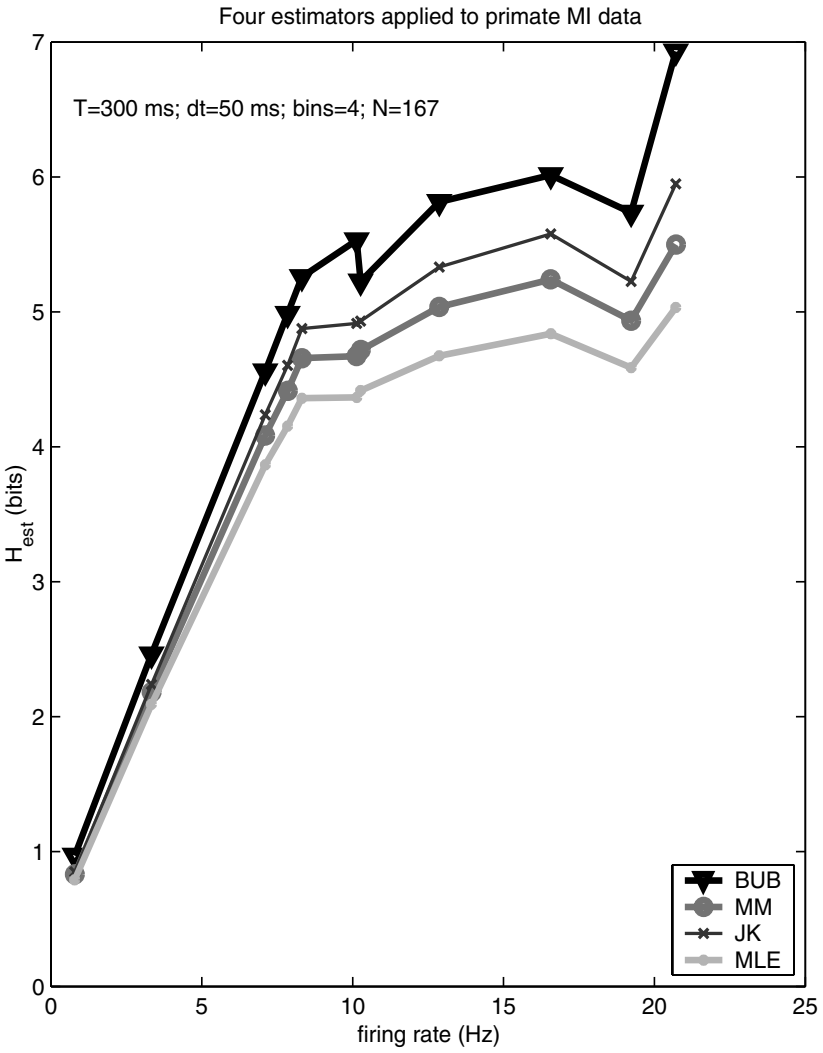

Figure 10: Estimated entropy of individual spike trains from 11 simultaneously recorded primate motor cortical neurons. Single units were recorded as a monkey performed a manual random tracking task (Paninski et al., 1999, 2003). N here refers to the number of trials (the stimulus was drawn i.i.d. on every trial). Each point on the x-axis represents a different cell; spike trains were 300 ms long, discretized into 50 ms bins of $0, 1, 2$, or > 2 spikes each; $m = 4^6$.

while a monkey moved its hand according to a stationary, two-dimensional, filtered gaussian noise process. We show results for 11 cells, simultaneously recorded during a single experiment, in Figure 10; note, however, that we are estimating the entropy of single-cell spike trains, not the full multicell spike train. (For more details on the experimental procedures, see Paninski, Fellows, Hatsopoulos, & Donoghue, 1999, 2003.)

With real data, it is, of course, impossible to determine the true value of H, and so the detailed error calculations performed above are not possible here. Nevertheless, the behavior of these estimators seems to follow the trends seen in the simulated data. We see the consistent slow increase in our estimate as we move from $\hat{H}_{MLE}$ to $\hat{H}_{MM}$ to $\hat{H}_{JK}$, and then a larger jump as we move to $\hat{H}_{BUB}$. This is true even though the relevant timescales (roughly defined as the correlation time of the stimulus) in the two experiments differed by about three orders of magnitude. Similar results were obtained for both the real and simulated data using a variety of other discretization parameters (data not shown). Thus, as far as can be determined, our conclusions about the behavior of these four estimators, obtained using the analytical and numerical techniques described above, seem to be consistent with results obtained using physiological data.

In all, we have that the new estimator performs quite well in a uniform sense. This good performance is especially striking in, but not limited to, the case when N/m is $O(1)$. We emphasize that even at the points where $H = \hat{H} = 0$ (and therefore the three most common estimators perform well, in a trivial sense), the new estimator performs reasonably; by construction, $\hat{H}_{BUB}$ never exhibits blowups in the expected error like those seen with $\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$. Given the fact that we can easily tune the bias of the new estimator at points where $H \approx 0$, by adjusting λ_0, $\hat{H}_{BUB}$ appears to be a robust and useful new estimator. We offer Matlab code, available on-line at http://www.cns.nyu.edu/∼liam, to compute the exact bias and Steele variance bound for any $\hat{H}_a$, at any distribution p, if the reader is interested in more detailed investigation of the properties of this class of estimator.

8 Directions for Future Work

We have left a few important open problems. Below, we give three somewhat freely defined directions for future work, along with a few preliminary results.

8.1 Bayes. All of our results here have been from a minimax, or "worst-case," point of view. As discussed above, this approach is natural if we know very little about the underlying probability measure. However, in many cases, we do know something about this underlying p. We might know that the spike count is distributed according to something like a Poisson distribution or that the responses of a neuron to a given set of stimuli can be

fairly well approximated by a simple dynamical model, such as an IF cell. How do we incorporate this kind of information in our estimates? The fields of parametric and Bayesian statistics address this issue explicitly. We have not systematically explored the parametric point of view—this would entail building a serious parametric model for spike trains and then efficiently estimating the entropy at each point in the parameter space—although this approach has been shown to be powerful in a few select cases. The Bayesian approach would involve choosing a suitable a priori distribution on spike trains and then computing the corresponding MAP or conditional mean estimator; this approach is obviously difficult as well, and we can give only a preliminary result here.

Wolpert and Wolf (1995) give an explicit formula for the Bayes' estimate of H and related statistics in the case of a uniform prior on the simplex. We note an interesting phenomenon relevant to this estimator: as m increases, the distribution on H induced by the flat measure on the simplex becomes concentrated around a single point, and therefore the corresponding Bayes' problem becomes trivial as $m \to \infty$, quite the opposite of the situation considered in the current work. (Nemenman, Shafee, & Bialek, 2002, independently obtained a few interesting results along these lines.) The result is interesting in its own right; its proof shares many of the features (concentration of measure and symmetry techniques) of our main results in the preceding sections.

More precisely, we consider a class of priors determined by the following "sort-difference" procedure: fix some probability measure P on the unit interval. Choose $m-1$ independent samples distributed according to P; sort the samples in ascending order, and call the sorted samples $\{x_i\}_{0<i<m}$. Define $q_1 = x_1$, $q_m = 1 - x_{m-1}$, and $q_i = x_i - x_{i-1}$ for all other i. This procedure therefore generates random probability measures q on m bins; in different language, the sort-difference procedure induces a prior on the m-simplex. (If P is the uniform density on the interval, for example, this prior is uniform on the m-simplex; this is the main case considered in Wolpert & Wolf, 1995.) The prior on q induces a prior on H, and this prior on H, in turn, happens to have a surprisingly small variance, for reasons quite similar to the reasons $\hat{H}$ has a surprisingly small variance: the entropy functional $H(p)$ is a symmetric and fairly smooth functional of p. So, let the prior on H, $P(H)$, be generated by this sort-difference procedure and assume for technical simplicity that the interval measure $P[0, 1]$ has a density component, p. We have the following crude but interesting result:

Theorem 6. *If p is bounded away from zero, then H is normally concentrated with rate $m^{1/3}$, that is, for fixed a,*

$$p(|H - E(H)| > a) = O(e^{-Cm^{1/3}a^2}),$$

for any constant $a > 0$ and some constant C.

In fact, it is possible to prove much more: the uniform measure on the simplex (and more generally, any prior induced by the sort-difference procedure, under some conditions on the interval measure P) turns out to induce an asymptotically normal prior on H, with variance decreasing in m. We can calculate the asymptotic mean of this distribution by using linearity of expectation and symmetry techniques like those used in section 5. In the following, assume for simplicity that P is equivalent to Lebesgue measure (that is, P is absolutely continuous with respect to Lebesgue measure, and vice versa); this is a technical condition that can be relaxed at the price of slightly more complicated formulas. We have the following:

Theorem 7. *P(H) is asymptotically normal, with*

$$Var(H) \sim \frac{1}{m}$$

and asymptotic mean calculated as follows.

Let q be the sorted, normalized density corresponding to a measure drawn according to the prior described above; define

$$F_p(v) \equiv \int_0^v du \int_0^1 dt p(t)^2 e^{-up(t)},$$

and

$$q'_\infty \equiv F_p^{-1},$$

where the inverse is taken in a distributional sense. Then

$$\|q - q'_\infty\|_1 \to 0$$

in probability and

$$E(H) \to h(q'_\infty) + \log(m),$$

where $h(.)$ denotes differential entropy.

F_p above is the cumulative distribution function of the p-mixture of exponentials with rate $p(t)^{-1}$ (just as $p'_{c,\infty}$ in theorem 3 was defined as the inverse cumulative distribution function (c.d.f) of a mixture of Poisson distributions). If P is uniform, for example, we have that $\|q - q'_\infty\|_1 \to 0$ in probability, where

$$q'_\infty(t) = -\log(t),$$

and

$$H(q) \to h(q'_\infty) + \log(m) = \log m + \int_0^1 dt \log(t) \log(-\log(t))$$

in probability.

8.2 Adaptive Partitioning. As emphasized in section 1, we have restricted our attention here to partitions, "sieves," S and T, which do not depend on the data. This is obviously a strong condition. Can we obtain any results without this assumption?

As a start, we have the following consistency result, stated in terms of the measure of the richness of a partition introduced by Vapnik and Chervonenkis (1971), $\Delta_N(\mathcal{A}_\mathcal{F})$ (the shatter coefficient of the set of allowed partitions, defined in the appendix; m is, as in the preceding, the maximal number of elements per partition, $\mathcal{F}$; Devroye et al., 1996):

Theorem 8. *If* $\log \Delta_N(\mathcal{A}_\mathcal{F}) = o\left(\frac{N}{(\log m)^2}\right)$ *and* $\mathcal{F}$ *generates* $\sigma_{x,y}$ *a.s.,* $\hat{I}$ *is consistent in probability;* $\hat{I}$ *is consistent a.s. under the slightly stronger condition*

$$\sum \Delta_N(\mathcal{A}_\mathcal{F}) e^{\frac{-N}{(\log m)^2}} < \infty.$$

Note the slower allowed rate of growth of m. In addition, the conditions of this theorem are typically harder to check than those of theorem 1. For example, it is easy to think of reasonable partitioning schemes that do not generate $\sigma_{x,y}$ a.s.: if the support of P is some measurable proper subset of $X \times Y$, this is an unreasonable condition. We can avoid this problem by rephrasing the condition in terms of $\sigma_{x,y}$ restricted to the support of P (this, in turn, requires placing some kind of topology on $X \times Y$, which should be natural enough in most problems).

What are the benefits? Intuitively, we should gain in efficiency: we are putting the partitions where they do the most good (Darbellay & Vajda, 1999). We also gain in applicability, since in practice, all partition schemes are data driven to some degree. The most important application of this result, however, is to the following question in learning theory: How do we choose the most informative partition? For example, given a spike train and some behaviorally relevant signal, what is the most efficient way to encode the information in the spike train about the stimulus? More concretely, all things being equal, does encoding temporal information, say, preserve more information about a given visual stimulus than encoding spike rate information? Conversely, does encoding the contrast of a scene, for example, preserve more information about a given neuron's activity than encoding color? Given m code words, how much information can we capture about what this neuron is telling us about the scene? (See, e.g., Victor, 2000b, for recent work along these lines.)

The formal analog to these kinds of questions is as follows (see Tishby, Pereira, & Bialek, 1999, and Gedeon, Parker, & Dimitrov, 2003, for slightly more general formulations). Let $\mathcal{F}$ and $\mathcal{G}$ be classes of "allowed" functions on the spaces $\mathcal{X}$ and $\mathcal{Y}$. For example, $\mathcal{F}$ and $\mathcal{G}$ could be classes of partitioning operators (corresponding to the discrete setup used here) or spaces of linear projections (corresponding to the information-maximization approach to independent component analysis.) Then, given N i.i.d. data pairs in $X \times Y$, we are trying to choose $f_N \in \mathcal{F}$ and $g_N \in \mathcal{G}$ in such a way that

$$I(f_N(x); g_N(y))$$

is maximized. This is where results like theorem 8 are useful; they allow us to place distribution-free bounds on

$$P\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}} I(f(x); g(y)) - \hat{I}(f_N(x); g_N(y)) > \epsilon\right), \tag{8.1}$$

that is, the probability that the set of code words that looks optimal given N samples is actually ϵ-close to optimal. Other (distribution-dependent) approaches to the asymptotics of quantities like equation 8.1 come from the the theory of empirical processes (see, e.g., van der Vaart & Wellner, 1996). More work in this direction will be necessary to rigorously answer the bias and variance problems associated with these "optimal coding" questions.

8.3 Smoothness and Other Functionals. We end with a slightly more abstract question. In the context of the sieve method analyzed here, are entropy and mutual information any harder to estimate than any other given functional of the probability distribution? Clearly, there is nothing special about H (and, by extension, I) in the case when m and p are fixed; here, classical methods lead to the usual $N^{-1/2}$ rates of convergence, with a prefactor that depends on only m and the differential properties of the functional H at p; the entire basic theory goes through if H is replaced by some other arbitrary (smooth) functional.

There are several reasons to suspect, however, that not all functionals are the same when m is allowed to vary with N. First, and most obvious, $\hat{H}$ is consistent when $m = o(N)$ but not when $m \sim N$; simple examples show that this is not true for all functionals of p (e.g., many linear functionals on m can be estimated given fewer than N samples, and this can be extended to weakly nonlinear functionals as well). Second, classical results from approximation theory indicate that smoothness plays an essential role in approximability; it is well known, for example, that the best rate in the bias polynomial approximation problem described in section 6 is essentially determined by a modulus of continuity of the function under question (Ditzian & Totik, 1987), and moduli of continuity pop up in apparently very different functional estimation contexts as well (Donoho & Liu, 1991; Jongbloed,

2000). Thus, it is reasonable to expect that the smoothness of H, especially as measured at the singular point near 0, should have a lot to do with the difficulty of the information estimation problem. Finally, basic results in learning theory (Devroye et al., 1996; van der Vaart & Wellner, 1996; Cucker & Smale, 2002) emphasize the strong connections between smoothness and various notions of learnability. For example, an application of theorem II.2.3 of Cucker and Smale (2002) gives the exact rate of decay of our L_2 objective function, equation 6.3, in terms of the spectral properties of the discrete differential operator D, expressed in the Bernstein polynomial basis; however, it is unclear at present whether this result can be extended to our final goal of a useful asymptotic theory for the upper L_∞ bound, equation 6.2.

A few of the questions we would like to answer more precisely are as follows. First, we would like to have the precise minimax rate of the information estimation problem; thus far, we have only been able to bound this rate between $m \sim o(N)$ (see theorem 1) and $m \sim N^{1+\alpha}$, $\alpha > 0$ (see theorem 4). Second, how close does $\hat{H}_{BUB}$ come to this minimax rate? Indeed, does this estimator require fewer than m samples to learn the entropy on m bins, as Figure 3 seems to indicate? Finally, how can all of this be generalized for other statistical functionals? Is there something like a single modulus of continuity that controls the difficulty of some large class of these functional estimation problems?

9 Conclusions

Several practical conclusions follow from the results presented here; we have good news and bad news. First, the bad news.

- Past work in which N/m was of order 1 or smaller was most likely contaminated by bias, even if the jackknife or Miller correction was used. This is particularly relevant for studies in which multiple binning schemes were compared to investigate, for example, the role of temporal information in the neural code. We emphasize for future studies that m and N must be provided for readers to have confidence in the results of entropy estimation.
- Error bars based on sample variance (or resampling techniques) give very bad confidence intervals if m and N are large. That is, confidence intervals based on the usual techniques do not contain the true value of H or I with high probability. Previous work in the literature often displays error bars that are probably misleadingly small. Confidence intervals should be of size

$$\sim B(\hat{H}, N/m) + N^{-1/2} \log(\min(m, N)),$$

where the bias term B can be calculated using techniques described in section 5.

Now the good news:

- This work has given us a much better understanding of exactly how difficult the information estimation problem is and what we can hope to accomplish using nonparametric techniques, given physiologically plausible sample sizes.
- We have obtained rigorous (and surprisingly general) results on bias, variance, and convergence of the most commonly employed estimators, including the best possible generalization of Miller's well-known $\frac{1}{N}$ bias rate result. Our analysis clarifies the relative importance of minimizing bias or variability depending on N and m, according to the bias-variance balance function introduced at the end of section 4.
- We have introduced a promising new estimator, one that comes equipped with built-in, rigorous confidence intervals. The techniques used to derive this estimator also lead to rigorous confidence intervals for a large class of other estimators (including the three most common $\hat{H}$).

Appendix A: Additional Results

A.1 Support. One would like to build an estimator that takes values strictly in some nice set around the true H, say, an interval containing H whose length shrinks as the number of samples, N, increases. This would give us strong "error bars" on our estimate of H; we would be absolutely certain that our estimate is close to the true H. The MLE for entropy has support on $[0, \log(\min(N, m))]$. A simple variational argument shows that any estimator, T, for H on m bins is inadmissible if T takes values outside $[0, \log m]$. Similarly, any estimator, T, for I on $m_S \times m_T$ bins is inadmissible if T takes values outside $[0, \log(\min(m_S, m_T))]$. It turns out that this is the best possible, in a sense: there do not exist any nontrivial estimators for entropy that are strictly greater or less than the unknown H. In fact, the following is true:

Proposition 7. *There is no estimator T and corresponding a, b, $0 < a <= 1$, $1 <= b < \infty$, such that the support of T is the interval $[aH, bH]$, for all values of the entropy, H.*

Proof. Suppose such an $a > 0$ exists. If so, $T(\omega)$ must be nonzero for all possible values of the data, ω (the data can be represented as an N-sequence of integers, $1 \leq \omega(i) \leq m$). But then there must exist some ω_0, with $p(\omega_0) > 0$, such that $T(\omega_0) > 0$. By choosing H such that $0 < H < T(\omega_0)$, we force a contradiction. The proof for b is similar.

A similar result obviously holds for mutual information.

A.2 Bias. It turns out that no unbiased estimators for H or I exist in the discrete setting. This fact seems to be known among information theorists, but we have not seen it stated in the literature. The proof is quite short, so we provide it here.

Proposition 8. *No unbiased estimator for entropy or mutual information exists.*

Proof. For any estimator T of the entropy of a multinomial distribution, we can write down the mean of T:

$$E(T) = \sum_{\omega \in \{1,\dots,m\}^N} P(\omega)T(\omega),$$

where $\{1, \dots, m\}^N$ is the sample space (i.e, each ω, as above, corresponds to an m-ary sequence of length N). Since ω_j is drawn i.i.d. from the discrete distribution p, $P(\omega)$ is given by

$$P(\omega) = \prod_{j=1}^{N} p_{\omega_j},$$

and so the mean of T is a polynomial function of the multinomial probabilities p_i. The entropy, on the other hand, is obviously a nonpolynomial function of the p_i. Hence, no unbiased estimator exists. The proof for I is identical.

The next (easy) proposition provides some more detail. The proof is similar to that of Proposition 7 and is therefore omitted.

Proposition 9. *(a) If T is a nonnegatively biased estimator for the entropy of a multinomial distribution on m bins, with $T(\omega) \in [0, \log(m)]$ $\forall \omega \in \{1, \dots, m\}^N$, then*

$$T(\omega) = \log(m) \ \ \forall \omega \in \{1, \dots, m\}^N.$$

(b) If T is a nonpositively biased estimator for the mutual information of a multinomial distribution on m_S, m_T bins, with $T(\omega) \in [0, \log(\min(m_S, m_T))]$, then

$$T(\omega) = 0 \ \ \forall \omega \in \Omega.$$

(c) If T is a nonnegatively biased estimator for the mutual information of a multinomial distribution on m_S, m_T bins, with $T(\omega) \in [0, \log(\min(m_S, m_T))]$,

then

$$T(\omega) = \log(\min(m_S, m_T)) \ \ \forall \omega \in \Omega.$$

A.3 Minimax Properties of σ-Symmetric Estimators. Let the error metric $D(T, \theta)$ be nice—convex in T, jointly continuous in T and θ, positive away from $T = \theta$, and bounded below. (The metrics given by

$$D(T, \theta) \equiv (T - \theta)^p, 1 \leq p < \infty$$

are good examples.) The following result partially justifies our focus throughout this article on estimators that are permutation symmetric (denoted σ-symmetric in the following).

Proposition 10. *If the error metric D is nice, then a σ-symmetric minimax estimator exists.*

Proof. Existence of a minimax estimator (see also Schervish, 1995): whenever $\max_\theta E_\theta(D)$ is a continuous function of the estimator T, a minimax estimator exists, since T can be taken to vary over a compact space (namely, $[0, \log m]^{m^N}$). But $\max \theta(E_\theta(D))$ is continuous in T whenever $E(D)$ is jointly continuous in T and θ. This is because $E(D)$ is uniformly continuous in θ and T, since, again, θ and T vary over compact spaces. $E(D)$ is jointly continuous in θ and T by the continuity of D and the fact that $E(D)$ is defined by a finite sum.

Existence of a symmetric minimax estimator: this is actually a special case of the Hunt-Stein theorem (Schervish, 1995). Any asymmetric minimax estimator, T, in the current setup achieves its maximum, $\max_\theta(E_\theta(D))$, by the arguments above. However, the corresponding symmetrized estimator, $T_\sigma(\omega) = (1/|\sigma|) \sum_\sigma T(\sigma(\omega))$, has expected error, which is less than or equal to $\max_\theta(E_\theta(D))$, as can be seen after a rearrangement and an application of Jensen's inequality. Therefore, T_σ is minimax (and obviously symmetric).

A.4 Insufficiency of Symmetric Estimators. The next result is perhaps surprising.

Proposition 11. *The MLE is not sufficient. In fact, the empirical histograms are minimal sufficient; thus, no σ-symmetric estimator is sufficient.*

Proof. A simple example suffices to prove the first statement. Choose as a prior on p:

$$P(p(1) = \epsilon; p(2) = 1 - \epsilon) = .5$$
$$P(p(1) = 0; p(2) = 1) = .5,$$

for some $\epsilon > 0$. For this P, $H(p) \to \hat{H} \to \{n_i\}$ does not form a Markov chain; the symmetry of $\hat{H}$ discards information about the true underlying H (namely, observation of a 1 tells us something very different than does observation of a 2). This property is clearly shared by any symmetric estimator.

The fact that the empirical histograms are minimal sufficient follows, for example, from Bahadur's theorem (Schervish, 1995), and the fact that the empirical histograms are complete sufficient statistics.

In other words, any σ-symmetric estimator necessarily discards information about H, even though H itself is σ-symmetric. This indicates the importance of priors; the nonparametric minimax approach taken here (focusing strictly on symmetric estimators for a large part of the work, as justified by proposition 10) should be considered only a first step. To be more concrete, in many applications, it is natural to guess that the underlying measure p has some continuity properties; therefore, estimators that take advantage of some underlying notion of continuity (e.g., by locally smoothing the observed distributions before estimating their entropy) should be expected to perform better (on average, according to this mostly continuous prior) than the best σ-symmetric estimator, which necessarily discards all topological structure in the underlying space $\mathcal{X}$. (See, e.g., Victor, 2002, for recent work along these lines.)

Appendix B: Proofs

We collect some deferred proofs here. To conserve space, we omit some of the easily verified details. The theorems are restated for convenience.

B.1 Consistency.

Statement (Theorem 1). *If* $m_{S,N} m_{T,N} = o(N)$ *and* σ_{S_N,T_N} *generates* $\sigma_{X,Y}$, *then* $\hat{I} \to I$ *a.s. as* $N \to \infty$.

Theorem 1 is a consequence of the following lemma:

Statement (Lemma 1). *If* $m = o(N)$, *then* $\hat{H} \to H_N$ *a.s.*

Proof. First, by the exponential bound of Antos and Kontoyiannis, expression 3.4, and the Borel-Cantelli lemma, $\hat{H}_N \to H_N$ *a.s.* if the (nonrandom) function $E(\hat{H}_N) \uparrow H_N$. This convergence in expectation is a consequence of the local expansion for the bias of the MLE, expression 4.2, and proposition 1 of section 4.

Proof of Theorem 1. First, some terminology: by $\hat{I} \to I$ *a.s.*, we mean that if $I = \infty$, $p((\hat{I}_N < c)$ *i.o.*$) = 0 \; \forall c < \infty$, and if $I < \infty$, $p((|\hat{I}_N - I| > \epsilon)$ *i.o.*$) = 0$

$\forall \epsilon > 0$. ("I.o." stands for "infinitely often.") In addition, we call the given σ-algebra on $X \times Y$ (the family of sets on which the probability measure $P(X, Y)$ is defined) $\sigma_{X,Y}$, and the sub-σ-algebra generated by S and T $\sigma_{S,T}$.

Now, the proof: it follows from Shannon's formula for mutual information in the discrete case that

$$|\hat{I}(S_N, T_N) - I(S_N, T_N)| \leq |\hat{H}(S) - H(S)| + |\hat{H}(T) - H(T)| + |\hat{H}(S, T) - H(S, T)|.$$

Thus, the lemma gives

$$\hat{I}_N \to I(S_N, T_N) \text{ a.s.}$$

whenever $m_S m_T / N \to 0$.

It remains only to show that the (nonrandom) function $I(S_N, T_N) \to I$; this follows from results in standard references, such as Billingsley (1965) and Kolmogorov (1993), if either

$$\sigma_{S_1,T_1} \subseteq \sigma_{S_2,T_2} \subseteq \cdots \subseteq \sigma_{S_N,T_N} \subseteq \cdots$$

and

$$\sigma_{X,Y} = \cup_N \sigma_{S_N,T_N},$$

or

$$\sup_{A \in \sigma_{S_N,T_N}, B \in \sigma_{X,Y}} \rho(A, B) \to 0,$$

where

$$\rho(A, B) \equiv P(AB^c \cup A^c B).$$

If either of these conditions holds, we say that σ_{S_N,T_N} generates $\sigma_{X,Y}$.

B.2 Central Limit Theorem.

Statement (Theorem 2). *Let*

$$\sigma_N^2 \equiv Var(-\log p_{T_N}) \equiv \sum_{i=1}^{m} p_{T_N,i}(-\log p_{T_N,i} - H_N)^2.$$

If $m_N \equiv m = o(N^{1/2})$, *and*

$$\liminf_{N \to \infty} N^{1-\alpha} \sigma_N^2 > 0$$

for some $\alpha > 0$, *then* $\left(\frac{N}{\sigma_N^2}\right)^{1/2} (\hat{H} - H_N)$ *is asymptotically standard normal.*

Proof. The basic tool, again, is the local expansion of H_{MLE}, expression 4.1. We must first show that the remainder term becomes negligible in probability on a $\sqrt{N}$ scale, that is,

$$\sqrt{N} D_{KL}(p_N; p) = o_p(1).$$

This follows from the formula for $E_p(D_{KL}(p_N; p))$, then Markov's inequality and the nonnegativity of D_{KL}.

So it remains only to show that $dH(p; p_N - p)$ is asymptotically normal. Here we apply a classical theorem on the asymptotic normality of double arrays of infinitesimal random variables:

Lemma. *Let $\{x_{j,N}\}, 1 \le N \le \infty, 1 \le j \le N$ be a double array of rowwise i.i.d. random variables with zero mean and variance σ_N^2, with distribution $p(x, N)$ and satisfying $\sigma_N^2 = 1/N$ for all N. Then $\sum_{j=1}^{N} x_{j,N}$ is asymptotically normal, with zero mean and unit variance, iff $\{x_{j,N}\}$ satisfy the Lindeberg (vanishing tail) condition: for all $\epsilon > 0$,*

$$\sum_{j=1}^{N} \int_{|x|>\epsilon} x^2 \, dp(x, N) = o(1). \tag{B.1}$$

The conditions of the theorem imply the Lindeberg condition, with $\{x_{j,n}\}$ replaced by $\frac{1}{\sqrt{N\sigma^2}}(dH(p; \delta_j - p) - H)$. To see this, note that the left-hand side of equation B.1 becomes, after the proper substitutions,

$$\frac{1}{\sigma^2} \sum_{p_j:\ (N\sigma^2)^{-\frac{1}{2}} |(\log p_j) - H| > \epsilon} p_j \log^2 p_j,$$

or

$$\frac{1}{\sigma^2} \left(\sum_{p_j:\ p_j > e^{\epsilon (N\sigma^2)^{\frac{1}{2}} + H}} p_j \log^2 p_j + \sum_{p_j:\ p_j < e^{H - \epsilon (N\sigma^2)^{\frac{1}{2}}}} p_j \log^2 p_j \right).$$

The number of terms in the sum on the left is less than or equal to

$$e^{-\epsilon (N\sigma^2)^{\frac{1}{2}} - H}.$$

Since the summands are bounded uniformly, this sum is $o(1)$. On the other hand, the sum on the right has at most m terms, so under the conditions of the theorem, this term must go to zero as well, and the proof is complete.

We have proven the above a.s. and $\sqrt{N}$ consistency theorems for $\hat{H}_{MLE}$ only; the extensions to $\hat{H}_{MM}$ and $\hat{H}_{JK}$ are easy and are therefore omitted.

B.3 Variance Bounds à la Steele. For the σ-symmetric statistic $H_a(\{x_j\}))$ $= \sum_j a_{j,N} h_{j,N}$, Steele's inequality reads:

$$\mathrm{Var}(H_a) \leq \frac{N}{2} E((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2),$$

where x'_N is a sample drawn independently from the same distribution as x_j. The linear form of H_a allows us to exactly compute the right-hand side of the above inequality. We condition on a given histogram, $\{n_i\}_{i=1,\ldots,m}$:

$$\begin{aligned} &E(H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2) \\ &\quad = \sum_{\{n_i\}} p(\{n_i\}) E((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2 | \{n_i\}). \end{aligned}$$

Now we rewrite the inner expectation on the right-hand side:

$$E((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2 \mid \{n_i\}) = E((D_- + D_+)^2 \mid \{n_i\}),$$

where

$$D_- \equiv a_{n_{x_N}-1,N} - a_{n_{x_N},N}$$

is the change in $\sum_j a_{j,N} h_{j,N}$ that occurs when a random sample is removed from the histogram $\{n_i\}$, according to the probability distribution $\{n_i\}/N$, and D_+ is the change in $\sum_j a_{j,N} h_{j,N}$ that occurs when a sample is randomly (and conditionally independently, given $\{n_i\}$) added back to the x_N-less histogram $\{n_i\}$, according to the true underlying measure p_i.

The necessary expectations are as follows. For $1 \leq j \leq N$, define

$$D_j \equiv a_{j-1} - a_j.$$

Then

$$\begin{aligned} E(D_-^2 \mid \{n_i\}) &= \sum_i \frac{n_i}{N} D_{n_i}^2, \\ E(D_+^2 \mid \{n_i\}) &= \sum_i p_i \left(\frac{n_i}{N} D_{n_i}^2 + \left(1 - \frac{n_i}{N}\right) D_{n_i+1}^2 \right), \end{aligned}$$

and

$$\begin{aligned} E(D_+ D_- \mid \{n_i\}) &= E(D_+ \mid \{n_i\}) E(D_- \mid \{n_i\}) \\ &= -\left(\sum_i \frac{n_i}{N} D_{n_i} \right) \left(\sum_i p_i \left(\frac{n_i}{N} D_{n_i} + \left(1 - \frac{n_i}{N}\right) D_{n_i+1} \right) \right). \end{aligned}$$

Taking expectations with respect to the multinomial measure $p(\{n_i\})$, we have

$$E(D_-^2) = \sum_{i,j} \frac{j}{N} D_j^2 B_j(p_i),$$

$$E(D_+^2) = \sum_{i,j} \left(\frac{j}{N} D_j^2 + \left(1 - \frac{j}{N}\right) D_{j+1}^2 \right) p_i B_j(p_i), \tag{B.2}$$

and

$$E(D_+ D_-) = -\sum_{i,i',j,k} \frac{j}{N} D_j \left(\frac{k}{N} D_k + \left(1 - \frac{k}{N}\right) D_{k+1} \right) p_{i'} B_{j,k}(p_i, p_{i'}),$$

where B_j and $B_{j,k}$ denote the binomial and trinomial polynomials, respectively:

$$B_j(t) \equiv \binom{N}{j} t^j (1-t)^{N-j};$$

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

The obtained bound,

$$\mathrm{Var}(H_a) \leq \frac{N}{2}(E(D_-^2) + 2E(D_- D_+) + E(D_+^2)),$$

may be computed in $O(N^2)$ time. For a more easily computable ($O(N)$) bound, note that $E(D_-^2) = E(D_+^2)$, and apply Cauchy-Schwartz to obtain

$$\mathrm{Var}(H_a) \leq 2NE(D_-^2).$$

Under the conditions of theorem 2, this simpler bound is asymptotically tight to within a factor of two. Proposition 6 is proven with devices identical to those used in obtaining the bias bounds of proposition 4 (note the similarity of equations 5.1 and B.2).

B.4 Convergence of Sorted Empirical Measures.

Statement (Theorem 3). *Let P be absolutely continuous with respect to Lebesgue measure on the interval $[0,1]$, and let $p = dP/dm$ be the corresponding density. Let S_N be the m-equipartition of $[0,1]$, p' denote the sorted empirical measure, and*

$N/m \to c,\ 0 < c < \infty$. *Then:*

a. $p' \overset{L_1, a.s.}{\to} p'_{c,\infty}$, *with* $\|p'_{c,\infty} - p\|_1 > 0$. *Here* $p'_{c,\infty}$ *is the monotonically decreasing step density with gaps between steps* j *and* $j+1$ *given by*

$$\int_0^1 dt e^{-cp(t)} \frac{(cp(t))^j}{j!}.$$

b. *Assume* p *is bounded. Then* $\hat{H} - H_N \to B_{c,\hat{H}}(p)$ *a.s., where* $B_{c,\hat{H}}(p)$ *is a deterministic function, nonconstant in* p. *For* $\hat{H} = \hat{H}_{MLE}$,

$$B_{c,\hat{H}}(p) = h(p') - h(p) < 0,$$

where $h(.)$ *denotes differential entropy.*

Proof. We will give only an outline. By $\overset{L_1, a.s.}{\to}$, we mean that $\|p - p_N\|_1 \to 0$ *a.s.* By McDiarmid's inequality, for any distribution q,

$$\|q - p_N\|_1 \to E(\|q - p_N\|_1) \text{ a.s.}$$

Therefore, convergence to some p' in probability in L_1 implies almost sure convergence in L_1. In addition, McDiarmid's inequality hints at the limiting form of the ordered histograms. We have

$$sort\left(\frac{n_i}{N}\right) \to E\left(sort\left(\frac{n_i}{N}\right)\right),\ a.s.\ \forall 1 \le i \le m.$$

Of course, this is not quite satisfactory, since both $sort(\frac{n_i}{N})$ and $E(sort(\frac{n_i}{N}))$ go to zero for most i.

Thus, we need only to prove convergence of $sort(\frac{n_i}{N})$ to p' in L_1 in probability. This follows by an examination of the histogram order statistics $h_{n,j}$. Recall that these $h_{n,j}$ completely determine p_n. In addition, the $h_{n,j}$ satisfy a law of large numbers:

$$\frac{1}{m} h_{n,j} \to \frac{1}{m} E(h_{n,j})\ \ \forall j \le k,$$

for any finite k. (This can be proven, for example, using McDiarmid's inequality.) Let us rewrite the above term more explicitly:

$$\frac{1}{m} E(h_{n,j}) = \frac{1}{m} \sum_{i=1}^{m} E(1(n_i = j)) = \frac{1}{m} \sum_{i=1}^{m} \binom{N}{j} p_i^j (1 - p_i)^{N-j}.$$

Now we can rewrite this sum as an integral:

$$\frac{1}{m} \sum_{i=1}^{m} \binom{N}{j} p_i^j (1 - p_i)^{N-j} = \int_0^1 dt \binom{N}{j} p_n(t)^j (1 - p_n(t))^{N-j}, \tag{B.3}$$

where

$$p_n(t) \equiv m p_{\max(i|i<t)}$$

is the discretized version of p. As the discretization becomes finer, the discretized version of p becomes close to p:

$$p_n \to p[\mu],$$

where $[\mu]$ denotes convergence in Lebesgue measure on the interval $[0, 1]$ (this can be seen using approximation in measure by continuous functions of the almost surely finite function p). Since $N/m \to c$ and the integrand in equation B.3 is bounded uniformly in N, we have by the dominated convergence theorem that

$$\int_0^1 dt \binom{N}{j} p_n(t)^j (1 - p_n(t))^{N-j} \to \int_0^1 dt \frac{c^j}{j!} p(t)^j e^{-cp(t)}. \tag{B.4}$$

From the convergence of $h_{n,j}$, it easily follows that $p_n \to p'$ in L_1 in probability, where p' is determined by $E(h_{n,j})$ in the obvious way (since $p_n \to p'$ except perhaps on a set of arbitrarily small p'-measure). Since $\lim \frac{h_{N,0}}{m} > \int_0^1 dt\, 1(p = 0)$, $\|p - p'\|_1$ is obviously bounded away from zero.

Regarding the final claim of the theorem, the convergence of the $\hat{H}$ to $E(\hat{H})$ follows by previous considerations. We need prove only that $h(p') < h(p)$. After some rearrangement, this is a consequence of Jensen's inequality.

B.5 Sum Inequalities. For $f(p) = 1/p$, we have the following chain of implications:

$$\begin{aligned} &\sup_p \left| \frac{g(p)}{p} \right| = c \\ &\quad \Rightarrow |g(p)| \le cp \\ &\Rightarrow \sum_{i=1}^{m} g(i) \le c \sum p = c. \end{aligned}$$

For the f described in section 6.1,

$$f(p) = \begin{cases} m & p < 1/m, \\ 1/p & p \ge 1/m, \end{cases}$$

we have

$$\sup_p |f(p)g(p)| = c \Rightarrow \sum_i g(i) \le 2c,$$

since

$$\sum_i g(i) = \sum_{i:\, im \geq 1} g(i) + \sum_{i:\, im < 1} g(i);$$

the first term is bounded by c, by the above, and the second by $cm\frac{1}{m} = c$.

This last inequality gives a proof of Proposition 2.

Statement (Proposition 2).

$$\max_p \sigma^2 \sim (\log m)^2.$$

Proof. We have $\max_p \sigma^2(p) = O(\log(m)^2)$: plug in $g = p\log(p)^2$ and take the maximum of fg on the interval. To see that in fact $\max_p \sigma^2(p) \sim (\log(m)^2)$, simply maximize $\sigma^2(p)$ on the central lines (see section 7).

B.6 Asymptotic Bias Rate.

Statement (Theorem 5). *If* $m > 1$, $N \min_i p_i \to \infty$, *then*

$$\lim \frac{N}{m-1} B(\hat{H}_{MLE}) = -\frac{1}{2}.$$

Proof. As stated above, the proof is an elaboration of the proof of theorem 10.3.1 of Devore and Lorentz (1993). We use a second-order expansion of the entropy function:

$$H(t) = H(x) + (t-x)H'(x) + (t-x)^2\left(\frac{1}{2}H''(x) + h_x(t-x)\right),$$

where h is a remainder term. Plugging in, we have

$$-t\log t = -x\log x + (t-x)(-1-\log x) + (t-x)^2\left(\frac{1}{2}\frac{-1}{x} + h_x(t-x)\right).$$

After some algebra,

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

After some more algebra (mostly recognizing the mean and variance formulae for the binomial distribution), we see that

$$\lim_N N(B_N(H)(x) - H(x)) = H''(x)\frac{x(1-x)}{2} + R_N(x),$$

where

$$R_N(x) \equiv \frac{1-x}{2} - N\sum_{j=0}^{N} B_{j,N}(x)\frac{j}{N}\log\frac{j}{Nx}.$$

The proof of theorem 10.3.1 in Devore and Lorentz (1993) proceeds by showing that $R_N(x) = o(1)$ for any fixed $x \in (0, 1)$. We need to show that $R_N(x) = o(1)$ uniformly for $x \in [x_N, 1 - x_N]$, where x_N is any sequence such that

$$Nx_N \to \infty.$$

This will prove the theorem, because the bias is the sum of

$$B_N(H)(x) - H(x)$$

at m points on this interval. The uniform estimate essentially follows from the delta method (somewhat like Miller and Madow's original proof, except in one dimension instead of m): use the fact that the sum in the definition of $R_N(x)$ converges to the expectation (with appropriate cutoffs) of the function $t \log \frac{t}{x}$ with respect to the gaussian distribution with mean x and variance $\frac{1}{N}x(1-x)$. We spend the rest of the proof justifying the above statement.

The sum in the definition of $R_N(x)$ is exactly the expectation of the function $t \log \frac{t}{x}$ with respect to the binomial(N, x) distribution (in a slight abuse of the usual notation, we mean a binomial random variable divided by N, that is, rescaled to have support on $[0, 1]$). The result follows if a second-order expansion for $t \log \frac{t}{x}$ at x converges at an $o(1/N)$ rate in $Bin_{N,x}$-expectation, that is, if

$$E_{Bin_{N,x}} N \left[t \log \frac{t}{x} - (t - x) - \frac{1}{2x}(t - x)^2 \right] \equiv E_{Bin_{N,x}} g_{N,x}(t) = o(1),$$

for $x \in [x_N, 1-x_N]$. Assume, wlog, that $x_N \to 0$; in addition, we will focus on the hardest case and assume $x_N = o(N^{-1/2})$. We break the above expectation into four parts:

$$E_{Bin_{N,x}} g_{N,x}(t) = \int_0^{ax_N} g dBin_{N,x} + \int_{ax_N}^{x_N} g dBin_{N,x} + \int_{x_N}^{b_N} g dBin_{N,x} + \int_{b_N}^{1} g dBin_{N,x},$$

where $0 < a < 1$ is a constant and b_N is a sequence we will specify below. We use Taylor's theorem to bound the integrands near x_N (this controls the middle two integrals) and use exponential inequalities to bound the binomial measures far from x_N (this controls the first and the last integrals). The inequalities are due to Chernoff (Devroye et al., 1996): let B be $Bin_{N,x}$, and let a, b, and x_N be as above. Then

$$P(B < ax_N) < e^{aNx_N - Nx_N - Nx_N a \log a} \tag{B.5}$$

$$P(B > b_N) < e^{Nb_N - Nx_N - Nb_N \log \frac{b_N}{x_N}}. \tag{B.6}$$

Simple calculus shows that

$$\max_{t\in[0,ax_N]} |g_{N,x}(t)| = g_{N,x}(0) = \frac{Nx}{2}.$$

We have that the first integral is $o(1)$ iff

$$aNx_N - Nx_N - Nx_N a \log a + \log(Nx_N) \to -\infty.$$

We rearrange:

$$\begin{aligned} &aNx_N - Nx_N - Nx_N a \log a + \log(Nx_N) \\ &\quad = Nx_N(a(1 - \log a) - 1) + \log(Nx_N). \end{aligned}$$

Since

$$a(1 - \log a) < 1, \forall a \in (0, 1),$$

the bound follows. Note that this is the point where the condition of the theorem enters; if Nx_N remains bounded, the application of the Chernoff inequality becomes useless and the theorem fails.

This takes care of the first integral. Taylor's bound suffices for the second integral:

$$\max_{t\in[ax_N,x_N]} |g_{N,x}(t)| < \left| \max_{u\in[ax_N,x_N]} \frac{(t-x)^3}{-6u^2} \right|,$$

from which we deduce

$$\left| \int_{ax_N}^{x_N} g dBin_{N,x} \right| < \frac{N((1-a)x_N)^4}{-6(ax_N)^2} = o(1),$$

by the assumption on x_N.

The last two integrals follow by similar methods once the sequence b_N is fixed. The third integral dies if b_N satisfies the following condition (derived, again, from Taylor's theorem):

$$\frac{N(b_N - x_N)^4}{-6x_N^2} = o(1),$$

or, equivalently,

$$b_N - x_N = o\left(\frac{x_N^{1/2}}{N^{1/4}}\right);$$

choose b_N as large as possible under this constraint, and use the second Chernoff inequality, to place an $o(1)$ bound on the last integral.

B.7 Bayes Concentration.

Statement (Theorem 6). *If p is bounded away from zero, then H is normally concentrated with rate $m^{1/3}$, that is, for fixed a,*

$$p(|H - E(H)| > a) = O(e^{-Cm^{1/3}a^2}),$$

for any constant $a > 0$ and some constant C.

Proof. We provide only a sketch. The idea is that H almost satisifies the bounded difference condition, in the following sense: there do exist points $x \in [0,1]^m$ such that

$$\sum_{i=1}^{m} (\Delta H(x_i))^2 > m\epsilon_m^2,$$

say, where

$$\Delta H(x_i) \equiv \max_{x_i, x_i'} |H(x_1, \ldots, x_i, \ldots x_m) - H(x_1, \ldots, x_i', \ldots x_m)|,$$

but the set of such x—call the set A—is of decreasing probability. If we modify H so that $H' = H$ on the complement of A and let $H' = E(H \mid p_i \in A^c)$ on A, that is,

$$H'(x) = \begin{cases} H(x) & x \in A^c, \\ \frac{1}{P(A^c)} \int_{A^c} P(x)H(x) & q \in A, \end{cases}$$

then we have that

$$P(|H' - E(H')| > a) < e^{-a^2(m\epsilon_m^2)^{-1}}$$

and

$$P(H' \neq H) = P(A).$$

We estimate $P(A)$ as follows:

$$\begin{aligned} P(A) &\leq \int_{[0,1]^m} 1(\max_i \Delta H(x_i) > \epsilon_m)\, d \prod_{i=1}^{m} p(x_i) \\ &\leq \int 1(\max_i (x_{i+2} - x_i) > \epsilon_m)\, dp^m(x_i) \\ &\sim \int dt e^{\log p - \epsilon_m p m}. \end{aligned} \tag{B.7}$$

The first inequality follows by replacing the L_2 norm in the bounded difference condition with an L_∞ norm; the second follows from some computation

and the smoothness of $H(x)$ with respect to changes in single x_i. The last approximation is based on an approximation in measure by nice functions argument similar to the one in the proof of theorem 3, along with the well-known asymptotic equivalence (up to constant factors), as $N \to \infty$, between the empirical process associated with a density p and the inhomogeneous Poisson process of rate Np.

We estimate $|E(H) - E(H')|$ with the following hacksaw:

$$\begin{aligned} |E(H) - E(H')| &= \left| \int_A p(x)(H(x) - H'(x)) + \int_{A^c} p(x)(H(x) - H'(x)) \right| \\ &= \left| \int_A p(x)(H(x) - H'(x)) + 0 \right| \\ &\leq P(A) \log m. \end{aligned}$$

If $p > c > 0$, the integral in equation B.7 is asymptotically less than $ce^{-m\epsilon_m c}$; the rate of the theorem is obtained by a crude optimization over ϵ_m.

The proof of the CLT in Theorem 7 follows upon combining previous results in this article with a few powerful older results. Again, to conserve space, we give only an outline. The asymptotic normality follows from McLeish's martingale CLT (Chow & Teicher, 1997) applied to the martingale $E(H \mid x_1, \ldots, x_i)$; the computation of the asymptotic mean follows by methods almost identical to those used in the proof of theorem 3 (sorting and linearity of expectation, effectively), and the asymptotic variance follows upon combining the formulas of Darling (1953) and Shao and Hahn (1995) with an approximation-in-measure argument similar, again, to that used to prove theorem 3. See also Wolpert and Wolf (1995) and Nemenman et al. (2002) for applications of Darling's formula to a similar problem.

B.8 Adaptive Partitioning.

Statement (Theorem 8). *If* $\log \Delta_N(\mathcal{A}_\mathcal{F}) = \mathrm{o}\left(\frac{\mathrm{N}}{(\log \mathrm{m})^2}\right)$ *and* $\mathcal{F}$ *generates* $\sigma_{x,y}$ *a.s.,* $\hat{I}$ *is consistent in probability;* $\hat{I}$ *is consistent a.s. under the slightly stronger condition*

$$\sum \Delta_N(\mathcal{A}_\mathcal{F}) \mathrm{e}^{\frac{-\mathrm{N}}{(\log \mathrm{m})^2}} < \infty.$$

The key inequality, unfortunately, requires some notation. We follow the terminology in Devroye et al. (1996), with a few obvious modifications. We take, as usual, $\{x_j\}$ as i.i.d random variables in some probability space $\Omega, \mathcal{G}, \mathrm{P}$. Let $\mathcal{F}$ be a collection of partitions of Ω, with $\mathcal{P}$ denoting a given partition. $2^\mathcal{P}$ denotes, as usual, the "power set" of a partition, the set of all sets that can be built up by unions of sets in $\mathcal{P}$. We introduce the class of

sets $\mathcal{A}_\mathcal{F}$, defined as the class of all sets obtained by taking unions of sets in a given partition, $\mathcal{P}$. In other words,

$$\mathcal{A}_\mathcal{F} \equiv \{A\colon\ A \in 2^\mathcal{P}, \mathcal{P} \in \mathcal{F}\}.$$

Finally, the Vapnik-Chervonenkis "shatter coefficient" of the class of sets $\mathcal{A}_\mathcal{F}$, $\Delta_N(\mathcal{A}_\mathcal{F})$, is defined as the number of sets that can be picked out of $\mathcal{A}_\mathcal{F}$ using N arbitrary points ω_j in Ω:

$$\Delta_N(\mathcal{A}_\mathcal{F}) \equiv \max_{\{\omega_j\}\in\Omega^N} |\{\omega_j\} \cap A\colon\ A \in \mathcal{A}|.$$

The rate of growth in N of $\Delta_N(\mathcal{A}_\mathcal{F})$ provides a powerful index of the richness of the family of partitions $\mathcal{A}_\mathcal{F}$, as the following theorem (a kind of uniform LLN) shows; p here denotes any probability measure and p_N, as usual, the empirical measure.

Theorem (Lugosi & Nobel, 1996). *Following the notation above, for any* $\epsilon > 0$,

$$P\left(\sup_{\mathcal{P}\in\mathcal{F}} \sum_{A\in\mathcal{P}} |p_N(A) - p(A)| > \epsilon\right) \leq 8\Delta_N(\mathcal{A}_\mathcal{F})\mathrm{e}^{-\mathrm{N}\epsilon^2/512}.$$

Thus, this theorem is useful if $\Delta_N(\mathcal{A}_\mathcal{F})$ does not grow too quickly with N. As it turns out, $\Delta_N(\mathcal{A}_\mathcal{F})$ grows at most polynomially in N under various easy-to-check conditions. Additionally, $\Delta_N(\mathcal{A}_\mathcal{F})$ can often be computed using straightforward combinatorial arguments, even when the number of distinct partitions in $\mathcal{F}$ may be uncountable. (See Devroye et al., 1996, for a collection of instructive examples.)

Proof. Theorem 8 is proven by a Borel-Cantelli argument, coupling the above VC inequality of Lugosi and Nobel with the following easy inequality, which states that the entropy functional H is "almost L_1 Lipshitz":

$$|H(p) - H(q)| \leq H_2(2\|p - q\|_1) + 2\|p - q\|_1 \log(m - 1),$$

where

$$H_2(x) \equiv -x\log(x) - (1 - x)\log(1 - x)$$

denotes the usual binary entropy function on $[0, 1]$. We leave the details to the reader.

Acknowledgments

Thanks to S. Schultz, R. Sussman, and E. Simoncelli for many interesting conversations; B. Lau and A. Reyes for their collaboration on collecting the in vitro data; M. Fellows, N. Hatsopoulos, and J. Donoghue for their collaboration on collecting the primate MI data; and I. Kontoyiannis, T. Gedeon,

and A. Dimitrov for detailed comments on previous drafts. This work was supported by a predoctoral fellowship from the Howard Hughes Medical Institute.

References

Antos, A., & Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms, 19*, 163–193.

Azuma, K. (1967). Weighted sums of certain dependent variables. *Tohoku Mathematical Journal, 3*, 357–367.

Basharin, G. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and Its Applications, 4*, 333–336.

Beirlant, J., Dudewicz, E., Gyorfi, L., & van der Meulen, E. (1997). Nonparametric entropy estimation: An overview. *International Journal of the Mathematical Statistics Sciences, 6*, 17–39.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., & Warland, D. (1991). Reading a neural code. *Science, 252*, 1854–1857.

Billingsley, P. (1965). *Ergodic theory and information*. New York: Wiley.

Buracas, G., Zador, A., DeWeese, M., & Albright, T. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron, 5*, 959–969.

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

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Annals of Mathematical Statistics, 23*, 493–509.

Chow, Y., & Teicher, H. (1997). *Probability theory*. New York: Springer-Verlag.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletins of the American Mathematical Society, 39*, 1–49.

Darbellay, G., & Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory, 45*, 1315–1321.

Darling, D. (1953). On a class of problems related to the random division of an interval. *Annals of Mathematical Statistics, 24*, 239–253.

Dembo, A., & Zeitouni, O. (1993). *Large deviations techniques and applications*. New York: Springer-Verlag.

Devore, R., & Lorentz, G. (1993). *Constructive approximation*. New York: Springer-Verlag.

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.

Ditzian, Z., & Totik, V. (1987). *Moduli of smoothness*. Berlin: Springer-Verlag.

Donoho, D., & Liu, R. (1991). Geometrizing rates of convergence. *Annals of Statistics, 19*, 633–701.

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics, 9*, 586–596.

Gedeon, T., Parker, A., & Dimitrov, A. (2003). *Information distortion and neural coding*. Manuscript submitted for publication.

Gibbs, A., & Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review, 70*, 419–436.

Grenander, U. (1981). *Abstract inference*. New York: Wiley.

Jongbloed, G. (2000). Minimax lower bounds and moduli of continuity. *Statistics and Probability Letters, 50*, 279–284.

Kolmogorov, A. (1993). *Information theory and the theory of algorithms*. Boston: Kluwer.

Kontoyiannis, I. (1997). Second-order noiseless source coding theorems. *IEEE Transactions Information Theory, 43*, 1339–1341.

Lugosi, A., & Nobel, A. B. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics, 24*, 687–706.

McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in combinatorics* (pp. 148–188). Cambridge: Cambridge University Press.

Miller, G. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology II-B* (pp. 95–100). Glencoe, IL: Free Press.

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing, 14*. Cambridge, MA: MIT Press.

Paninski, L., Fellows, M., Hatsopoulos, N., & Donoghue, J. (1999). Coding dynamic variables in populations of motor cortex neurons. *Society for Neuroscience Abstracts, 25*, 665.9.

Paninski, L., Fellows, M., Hatsopoulos, N., & Donoghue, J. (2003). *Temporal tuning properties for hand position and velocity in motor cortical neurons*. Manuscript submitted for publication.

Paninski, L., Lau, B., & Reyes, A. (in press). Noise-driven adaptation: In vitro and mathematical analysis. *Neurocomputing*. Available on-line: http://www.cns.nyu.edu/ ∼liam/adapt.html.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems, 7*, 87–107.

Prakasa Rao, B. (2001). Cramer-Rao type integral inequalities for general loss functions. *TEST, 10*, 105–120.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical recipes in C*. Cambridge: Cambridge University Press.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Ritov, Y., & Bickel, P. (1990). Achieving information bounds in non- and semiparametric models. *Annals of Statistics, 18*, 925–938.

Schervish, M. (1995). *Theory of statistics*. New York: Springer-Verlag.

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shao, Y., & Hahn, M. (1995). Limit theorems for the logarithm of sample spacings. *Statistics and Probability Letters, 24*, 121–132.

Steele, J. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics, 14*, 753–758.

Strong, S. Koberle, R., de Ruyter van Steveninck R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters, 80*, 197–202.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In B. Hajek & R. S. Sreenivas (Eds.), *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*. Urbana, IL: University of Illinois Press.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation, 7*, 399–407.

van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. New York: Springer-Verlag.

Vapnik, V. N., & Chervonenkis, A. J. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications, 16*, 264–280.

Victor, J. (2000a). Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Computation, 12*, 2797–2804.

Victor, J. (2000b). How the brain uses time to represent and process visual information. *Brain Research, 886*, 33–46.

Victor, J. (2002). Binless strategies for estimation of information from neural data. *Physical Review E, 66*, 51903–51918.

Watson, G. (1980). *Approximation theory and numerical methods*. New York: Wiley.

Weaver, W., & Shannon, C. E. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Wolpert, D., & Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E, 52*, 6841–6854.

Received June 3, 2002; accepted Nov. 27, 2002.