Undersmoothed kernel entropy estimators

Liam Paninski and Masanao Yajima
Department of Statistics, Columbia University
liam@stat.columbia.edu, my2167@columbia.edu
http://www.stat.columbia.edu/∼liam

***Abstract*— We develop a "plug-in" kernel estimator for the differential entropy that is consistent even if the kernel width tends to zero as quickly as $1/N$, where N is the number of i.i.d. samples. Thus, accurate density estimates are not required for accurate kernel entropy estimates; in fact, it is a good idea when estimating entropy to sacrifice some accuracy in the quality of the corresponding density estimate.**

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

INTRODUCTION

The estimation of the entropy and of related quantities (mutual information, Kullback-Leibler divergence, etc.) from i.i.d. samples is a very well-studied problem. Work on estimating the discrete entropy began shortly after the appearance of Shannon's original work (Miller, 1955; Basharin, 1959; Antos and Kontoyiannis, 2001; Paninski, 2003). A variety of nonparametric approaches for estimating the differential entropy have been studied, including histogram-based estimators, "plug-in" kernel estimators, resampled kernel estimators, and nearest-neighbor estimators; see (Beirlant et al., 1997) for a nice review.

In particular, this previous work has established the consistency of several kernel- or nearest-neighbor-based estimators of the differential entropy, under certain smoothness or tail conditions on the underlying (unknown) distribution p. In the kernel case, consistency is established under the assumption that the kernel width scales more slowly than $1/N$ (Beirlant et al., 1997); this is the usual assumption guaranteeing that the corresponding kernel density estimate is consistent (not "undersmoothed"). While these consistency results are well-understood, worst-case error bounds — i.e., bounds on the estimator's average error over a large class of underlying probability measures p — are more rare.

Our main result here is an adaptation of the discrete (histogram-based) techniques of (Paninski, 2003; Paninski, 2004) to the kernel estimator case. This earlier work established universal consistency for a histogram-based estimator of the entropy assuming that the number of histogram bins, $m = m_N$, obeyed the scaling $m_N = O(N)$; in addition, nonparametric error bounds were established for any (m, N) pair. To adapt these results here we decompose the error of the kernel estimator into three parts: a (deterministic) smoothing error, and an estimation error consisting of the usual bias and variance terms. Smoothing error generically decreases with kernel width, and therefore it is beneficial to make the kernel width as small as possible; on the other hand, in the classical plug-in entropy estimators, making the kernel width too small can make the estimation error component (the bias plus the variance) large. We provide an estimator whose estimation error term may be bounded by a term which goes to zero even if the kernel width scales as $1/N$. Thus, accurate density estimates are not required for accurate kernel entropy estimates; in fact, it is a good idea when estimating entropy to sacrifice some accuracy in the quality of the corresponding density estimate (i.e., to undersmooth). Some comparisons on simulated data are provided.

LP was partially supported by an NSF CAREER award.

MAIN RESULTS

We assume that data $\{x_j\}$, $1 \le j \le N$, are drawn i.i.d. from some arbitrary probability measure p. We are interested in estimating the differential entropy of p (Cover and Thomas, 1991),

$$H(p) = \int -\frac{dp(s)}{ds} \log \frac{dp(s)}{ds} ds$$

(for clarity, we will restrict our attention here to the case that the base measure ds is Lebesgue measure on a finite one-dimensional interval $\mathcal{X}$ of length $\mu(\mathcal{X})$, though extensions of the following results to more general measure spaces are possible.)

We will consider kernel entropy estimators of the following form:

$$\hat{H} = \int g(\hat{p}(s)) ds,$$

where we define the kernel density estimate

$$\hat{p}(s) = \frac{1}{N} \sum_{j=1}^{N} k(s - x_j),$$

with $k(.)$ the kernel; as usual, $\int k ds = 1$ and $k \ge 0$. The standard "plug-in" estimator for the entropy is obtained by setting

$$g(u) = h(u) \equiv -u \log u;$$

our basic plan is to optimize $g(.)$, in some sense, to obtain a better estimate than the plug-in estimate.

Our development begins with the standard bias-variance decomposition for the squared error of the estimator:

$$E(H - \hat{H})^2 = (E_{app} + B(\hat{H}))^2 + V(\hat{H}),$$

with the approximation error

$$E_{app} = H(p * k) - H(p),$$

and the bias term

$$B(\hat{H}) = E_p(\hat{H}) - H(p * k)$$

defined relative to the smoothed measure

$$p * k(s) = E_p(\hat{p}(s)) = \int k(s - x) dp(x).$$

Note that E_{app} is generically positive and increasing with the kernel width (since smoothing tends to increase entropy), while the bias $B(\hat{H})$ of the standard plug-in estimator ($g(.) = h(.)$) is always negative, by Jensen's inequality.

Clearly, it is impossible to obtain any nontrivial risk bounds on the expected mean-square error of any estimator of the differential entropy, since we might have $H = -\infty$ (in the case that p is singular). Thus, instead of trying to obtain bounds on the full error $E(H(p) - \hat{H})^2$, our goal will be to bound the estimation error

$$E(H(p * k) - \hat{H})^2 = B(\hat{H})^2 + V(\hat{H}),$$

and then choose the kernel k so that the smoothing error E_{app} is as small as possible, under the constraint that the worst-case expected estimation error is acceptably small.

For this class of kernel entropy estimators, we have some simple bounds on the bias and variance (adapted from bounds derived in (Antos and Kontoyiannis, 2001; Paninski, 2003)). We may bound the variance $V(\hat{H}_{g,N})$ using McDiarmid's technique (Devroye et al., 1996; McDiarmid, 1989):

Lemma 1 (Variance bound, general kernel)**.**

$$V(\hat{H}_{g,N}) \le N \left(\int \sup_y \left| g(y) - g\left(y + \frac{k(s)}{N} \right) \right| ds \right)^2.$$

In the special case that $g(.)$ is Lipschitz, $\sup_{s,t}|g(s)-g(s+t)| \leq c|t|$, for some $0<c<\infty$, the bound simplifies considerably:

$$V(\hat{H}_{g,N}) \leq c^2/N.$$

Proof: McDiarmid's variance inequality (Devroye et al., 1996; McDiarmid, 1989) says that if we may bound the maximal coordinatewise difference

$$\sup_{x_1,\ldots,x_j,x_j',\ldots,x_N} \left| \hat{H}_{g,N}(x_1,\ldots,x_j,\ldots,x_N) - \hat{H}_{g,N}(x_1,\ldots,x_j',\ldots,x_N) \right| \leq c_j,$$

where $H_{g,N}(x_1,\ldots,x_N)$ denotes the estimator evaluated on some arbitrary configuration of the observed samples $\{x_j\}_{1\leq j\leq N}$, then

$$V(\hat{H}_{g,N}) \leq \frac{1}{4}\sum_j c_j^2.$$

We have here that c_j, as defined above, may be chosen as

$$c_j = 2\int \sup_y \left| g(y) - g\left(y+\frac{k(s)}{N}\right)\right| ds;$$

plugging in, we obtain the the general bound in the lemma.

In the Lipschitz case,

$$N\left(\int \sup_y \left| g(y) - g\left(y+\frac{k(s)}{N}\right)\right| ds\right)^2 \leq N\left(\int \frac{c}{N}k(s)ds\right)^2 = N(c/N)^2 = c^2/N,$$

where the first inequality follows by the Lipschitz condition and the first equality by the fact that the kernel k integrates to one.

Exponential tail bounds are also available (McDiarmid, 1989; Devroye et al., 1996; Antos and Kontoyiannis, 2001; Paninski, 2003) in case almost-sure results are desired, but these bounds will not be necessary here.

We now specialize to the simplest possible kernel, the step kernel of width w:

$$k_w(s) = \frac{1}{w}1\big(s\in[-w/2,w/2]\big).$$

In this case we only need to define $g(u)$ at the $N+1$ points $u = \{0,\frac{1}{Nw},\frac{2}{Nw},\ldots,\frac{1}{w}\}$, and we have the following simplification of Lemma 1:

Lemma 2 (Variance bound, step kernel)**.**

$$\sup_p V(\hat{H}_{g,N}) \leq Nw^2 \max_{0\leq j<N}\left[g\left(\frac{j+1}{Nw}\right)-g\left(\frac{j}{Nw}\right)\right]^2.$$

Proof: In this case it is easy to see that $\int \sup_y \left|g(y)-g\left(y+\frac{k_w(s)}{N}\right)\right| ds$ is bounded above by

$$w\max_{0\leq j<N}\left|g\left(\frac{j+1}{Nw}\right)-g\left(\frac{j}{Nw}\right)\right|;$$

the result now follows directly from Lemma 1.

We may compute the bias $B(\hat{H}_{g,N})$ exactly in this special step-kernel case:

$$\begin{aligned} B(\hat{H}_{g,N}) &= E_p(\hat{H}_{g,N}) - H(p*k_w) \\ &= -\int\left(h[p*k_w(s)] - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}[wp*k_w(s)]\right)ds, \quad (1)\end{aligned}$$

where we have abbreviated the binomial functions

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

the derivation of this formula exactly follows that in the discrete case, as described in (Paninski, 2003) (all that is required is an interchange of an integral and a finite sum). From this we may easily derive the following approximation-theoretic bound:

Lemma 3 (Bias bound)**.** [1]

$$\sup_p |B(\hat{H}_{g,N})| \leq \mu(\mathcal{X}) \max_{0\leq u\leq 1}\left|\frac{1}{w}h(u)+\log w - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}(u)\right|.$$

Proof: We apply the simple inequality $|\int_{\mathcal{X}} f(x)d\mu(x)| \leq \mu(\mathcal{X})\sup_x|f(x)|$ to the expression for the bias in equation (1). First we rewrite

$$\int h[p*k_w(x)]dx = \int h[\frac{1}{w}wp*k_w(x)]dx = \log w + \frac{1}{w}\int h[wp*k_w(x)]dx.$$

Now

$$B(\hat{H}_{g,N}) = -\int\left(\log w + \frac{1}{w}h[wp*k_w(x)] - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}[wp*k_w(x)]\right)dx,$$

so $|B(\hat{H}_{g,N})|$ is bounded above by

$$\begin{aligned} &\mu(\mathcal{X})\sup_x\left|\log w + \frac{1}{w}h[wp*k_w(x)] - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}[wp*k_w(x)]\right| \\ =\ &\mu(\mathcal{X})\max_{0\leq u\leq 1}\left|\log w + \frac{1}{w}h(u) - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}(u)\right|,\end{aligned}$$

since $0 \leq wp*k_w(x) \leq 1$. The maximum is obtained, by compactness and continuity of $h(u)$ and $B_{j,N}(u)$.

Note that each of the above bounds is distribution-free, that is, uniform over all possible underlying distributions p. We may combine these to obtain uniform bounds on the mean-square error:

$$\begin{aligned} &\sup_p E(\hat{H}_{g,N} - H(p*k_w))^2 \\ =\ &\sup_p\left[B(\hat{H}_{g,N})^2 + V(\hat{H}_{g,N})\right] \\ \leq\ &\left(\sup_p|B(\hat{H}_{g,N})|\right)^2 + \sup_p V(\hat{H}_{g,N}) \\ \leq\ &\mu(\mathcal{X})^2\max_{0\leq u\leq 1}\left|\frac{1}{w}h(u)+\log w - \sum_{j=0}^N g(\frac{j}{Nw})B_{j,N}(u)\right|^2 \\ &+Nw^2\max_{0\leq j<N}\left(g(\frac{j+1}{Nw})-g(\frac{j}{Nw})\right)^2 \\ =\ &\left(\frac{\mu(\mathcal{X})}{w}\right)^2\max_{0\leq u\leq 1}\left|h(u)+w\log w - \sum_{j=0}^N wg(\frac{j}{Nw})B_{j,N}(u)\right|^2 \\ &+N\max_{0\leq j<N}\left(wg(\frac{j+1}{Nw})-wg(\frac{j}{Nw})\right)^2. \quad (2)\end{aligned}$$

If we define

$$a(j/N) = w\left[g(j/N) - \log w\right],$$

then expression (2) simplifies to

$$\left(\frac{\mu(\mathcal{X})}{w}\right)^2\max_{0\leq u\leq 1}\left|h(u) - \sum_{j=0}^N a(\frac{j}{N})B_{j,N}(u)\right|^2 + N\max_{0\leq j<N}\left(a(\frac{j+1}{N})-a(\frac{j}{N})\right)^2.$$

In (Paninski, 2004) we proved that there exists a sequence of functions a_N (defined implicitly as the solution to a certain

[1] A direct generalization to the infinite $\mu(\mathcal{X})$ case is not possible without some restrictions on the decay of p. We will not pursue such bounds here.

approximation-theoretic convex optimization problem) such that

$$m_N^2 \max_{0\le u\le 1}\left|h(u)-\sum_{j=0}^{N} a_N(\frac{j}{N})B_{j,N}(u)\right|^2 + N\max_{0\le j<N}\left(a_N(\frac{j+1}{N})-a_N(\frac{j}{N})\right)^2$$

converges to zero as $N\to\infty$ for any sequence m_N satisfying $m_N = O(N)$.[2]

Now, setting $m_N = \mu(\mathcal{X})/w_N$, we may now easily deduce the main result of this paper:

Theorem 4. *Let $Nw_N \ge c > 0$, uniformly in N. There exists an estimator $\hat{H}_{g,N}$ for the entropy H which is uniformly smoothed-consistent in mean square; that is,*

$$\sup_p E(\hat{H}_{g,N} - H(p * k_{w_N}))^2 < \epsilon(c,N),$$

with $\epsilon(c,N)\searrow 0$ as $N\to\infty$, and the supremum is taken over all probability measures p.

Proof: We need only apply the main result of (Paninski, 2004) guaranteeing the existence of the sequence a_N described above, and then take $g(j/N) = \frac{1}{w}a_N(j/N) + \log w$.

As a corollary, it is easy to show that a uniformly consistent estimator exists if $Nw_N \to 0$ sufficiently slowly; as in (Paninski, 2004), this follows by a straightforward diagonalization argument. Note that $w_N = O(N^{-1})$ (and certainly $w_N = o(N^{-1})$) does not lead to consistent density estimates, even under smoothness restrictions on p (Paninski, 2003; Braess and Dette, 2004; Paninski, 2005). Thus the content of the theorem is that we can undersmooth the density and still estimate entropy well. In fact, undersmoothing is a good idea because it generically decreases the approximation bias E_{app}.

Finally, it is worth noting that an identical result may be obtained in the multidimensional case; the only difference in the statement and proof of the result is that in the general case the inverse measure of the support of our step kernel must be $O(N)$, whereas in the one-dimensional case (theorem 4) we restrict the inverse *length* w_N to be $O(N)$.

Numerical results

Sample-spacing estimators also have the "undersmoothing" property — consistent density estimates are not required for consistent entropy estimates (Beirlant et al., 1997). Thus it makes sense to compare the performance of the estimator introduced here with that of these sample-spacing estimators.

The m-sample spacing estimator is defined as follows. Given N real-valued samples X_i, we may form the usual order statistics $X_{(i)}$. The gaps between the i-th and $(i+m)$-th order statistics, $X_{(i+m)} - X_{(i)}$, are called the m-spacings. It is easy to form a density estimator based on these m-spacings (Beirlant et al., 1997), and plugging this estimator into the differential entropy formula (and performing a bias correction) gives the following estimator for the entropy:

$$\hat{H}^{(m)} \equiv \frac{1}{N}\sum_{i=1}^{N-m}\log\left(\frac{N}{m}(X_{(i+m)} - X_{(i)})\right) - \psi(m) + \log m,$$

where we have abbreviated the digamma function

$$\psi(x) = \left.\frac{\partial \log\Gamma(t)}{\partial t}\right|_{t=x}.$$

[2] In (Paninski, 2004) m was defined as the finite number of points on which the discrete probability measure was supported. Note that this definition of m is consistent with the definition of m in the case of a histogram-based method, in which we divide the space $\mathcal{X}$ into m bins and the effective kernel width w is exactly $\mu(\mathcal{X})/m$.

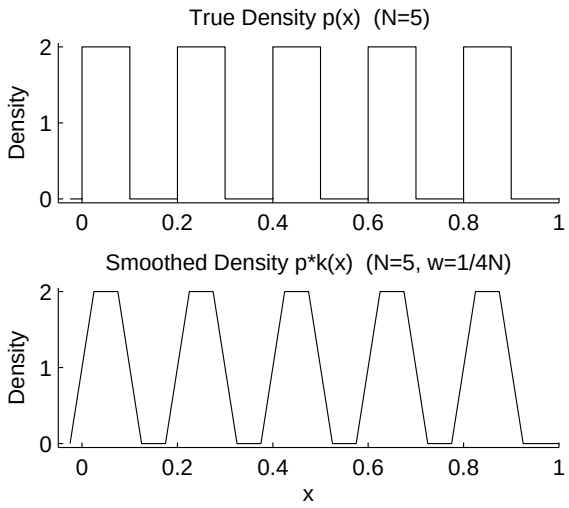

Fig. 1. **Top**: True density used for the simulations described in the text and in Fig. 2. **Bottom**: Smoothed density.

To compare the performance of the estimators, it is useful to choose a bounded, absolutely continuous density whose entropy is very vulnerable to oversmoothing, that is, a density p for which $H(p)$ and $H(p*k)$ are very different and therefore the approximation error E_{app} is large. One such density is of the one-dimensional sawtooth form

$$p(x) = p_N(x) \equiv \begin{cases} 2 & \frac{n-1}{N} \le x \le \frac{2n-1}{2N} \\ 0 & \frac{2n-1}{2N} \le x \le \frac{n}{N}, \end{cases}$$

where $n = 1, 2, \cdots, N$ (Fig. 1). (More generally, any density with large fluctuations on a $1/w$ scale will induce a large approximation error E_{app}; the density p chosen here just has a particularly convenient form.) The entropy $H(p)$ of this distribution can easily be calculated as $-\log(2)$.

The smoothed entropy $H(p*k)$ may also be computed explicitly here. The density $p*k$ is simply a sum of trapezoids, of the form

$$p{*}k(x) = \begin{cases} \frac{2}{w}(x - (\frac{n-1}{N} - \frac{w}{2})) & \frac{n-1}{N} - \frac{w}{2} \le x \le \frac{n-1}{N} + \frac{w}{2} \\ 2 & \frac{n-1}{N} + \frac{w}{2} \le x \le \frac{2n-1}{2N} - \frac{w}{2} \\ 2[1 - \frac{1}{w}(x - (\frac{2n-1}{2N} - \frac{w}{2}))] & \frac{2n-1}{2N} - \frac{w}{2} \le x \le \frac{2n-1}{2N} + \frac{w}{2} \\ 0 & \frac{2n-1}{2N} + \frac{w}{2} \le x \le \frac{n}{N}, \end{cases}$$

where we have assumed that $w < 1/N$. Thus for the smoothed entropy we obtain

$$\begin{aligned} H(p*k) &= -\int_{\Re} p*k(x)\log p*k(x)dx \\ &= -N\int_{\frac{w}{2}}^{\frac{1}{2N}-\frac{w}{2}} 2\log 2dx - 2N\int_0^w \frac{2x}{w}\log\frac{2x}{w}dx\} \\ &= -N\left(\frac{1}{2N} - w\right)2\log 2 - 2N\frac{w}{2}\int_0^2 y\log y dy \\ &= -N\left(\frac{1}{2N} - w\right)2\log 2 - \frac{Nw}{2}\left(y^2\log y - \frac{1}{2}y^2\right)\Big|_0^2 \\ &= -\log 2 + Nw. \end{aligned}$$

We illustrate the performance of the new kernel estimator (which we will refer to by the initials "BUB," for "best upper bound," as in (Paninski, 2003)) versus the m-spacing estimator with $m = 1$ (this value of m led to the best performance here; data not shown)

in Fig. 2.[3] The idea was to choose w to be as small as possible (to make the smoothing error $H(p*k) - H(p)$ as small as possible), within the constraint that the maximal error $\max_p[\hat{H} - H(p*k)]^2$ is decreasing as a function of N (to ensure smoothed-consistency of the estimate $\hat{H}$). This behavior is illustrated in Fig. 2: we see that the error bound does in fact tend to zero (albeit slowly), implying that $\hat{H} \to H(p*k)$ in mean square; at the same time, since $nW_N \to 0$, $H(p*k) \to H(p)$, and we have that $\hat{H}$ is not only smoothed-consistent in mean-square but in fact mean-square consistent for $H(p)$. On the other hand the m-spacing estimator has an asymptotic bias; since the m-spacing estimator is constructed from a density estimate whose kernel width, roughly speaking, would correspond to $1/[Np(x)]$, this estimator cannot detect the structure on the $o(1/N)$ scale which is necessary to consistently estimate $H(p)$ here. (But note that the m-spacing estimator can be superior in the case of an unbounded density p, where the smoothing error $H(p*k) - H(p)$ of the kernel estimator is large but where the small effective width $1/[Np(x)]$ of the m-spacing estimator can lead to a much smaller bias; data not shown.)

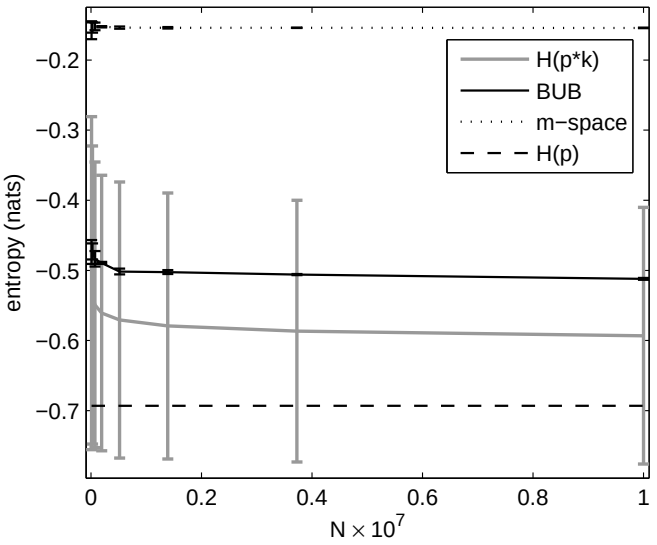

Fig. 2. Comparison of the performance of the m-spacing ($m = 1$) and BUB estimators applied to the density p shown in Fig. 1. For each of several values of the sample size N, we chose N i.i.d. samples from p (with N in the definition of p chosen to equal the sample size in each case; i.e., the number of sawtooths in the definition of p increases linearly with N), then replicated the experiment ten times, in order to obtain reliable estimates of the sample mean and standard deviation of the two estimates. This sample mean, plus and minus a single standard deviation, is plotted for the m-spacing and BUB estimates (dotted and solid black traces, respectively). Note the large positive asymptotic bias of the m-spacing estimator (the variance of both the m-spacing and BUB estimators are relatively negligible). The true value of the entropy, $H(p) = -\log 2$, is indicated by the dashed line; the gray trace shows the true smoothed entropy, plus or minus the square root of the maximal mean-squared error of the BUB estimator. Note that this maximal error tends to zero as $N \to \infty$, as does $H(p*k) \to H(p)$, for the values of w_N chosen here, implying mean-square consistency of $\hat{H}$ for $H(p)$.

CONCLUSIONS

We have presented a kernel density estimator of the entropy (based on a simple step kernel) which can be applied even when the kernel undersmooths the true underlying density (that is, when the kernel width tends to zero as quickly as $1/N$). This kernel estimator is shown to have better numerical performance than the classical m-spacing estimators when the underlying density p is very jagged. We anticipate that this new estimator will be useful in applications that require the estimation of differential entropy of a random vector, or of the mutual information between two random variables.

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

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., and van der Meulen, E. (1997). Nonparametric entropy estimation: an overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39.

Braess, D. and Dette, H. (2004). The asymptotic minimax risk for the estimation of constrained binomial and multinomial probabilities. *Sankhya*, 66:707–732.

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

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

[3]Kernel density estimators are typically computationally expensive, requiring $O(Nt)$ time to compute, where t denotes the number of points at which we evaluate the integrand in the definition of the entropy estimate; the m-spacing estimates, on the other hand, may be computed after a simple sorting operation which requires $O(N \log N)$ time (typically t is taken to be significantly larger than $\log N$; i.e., the m-spacing estimator is computationally cheaper). However, in the case of the step kernel used here, applied to one-dimensional data, it is possible to compute the density estimate, and therefore $\hat{H}$, in $O(N \log N)$ time: we need only sort the sample points (as in the case of the m-spacing estimator), then to compute the integral in the definition of $\hat{H}$ we need only keep track of the $2N$ points at which the density estimate $\sum_i k(x_i)$ jumps up or down (at the points $\{x_i - w/2\}$ and $\{x_i + w/2\}$, respectively); the whole computation requires just a couple lines of code.

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

Miller, G. (1955). Note on the bias of information estimates. In *Information theory in psychology II-B*, pages 95–100.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253.

Paninski, L. (2004). Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50:2200–2203.

Paninski, L. (2005). Variational minimax estimation of discrete distributions under KL loss. *Advances in Neural Information Processing Systems*, 17.