Ensembles of Nonparametric Entropy Estimators

Vy Nguyen, Majeed Thaika
andrewID: vyn, mthaika
36/10-702: Final Report

1 Introduction

Information entropy is the average amount of information produced by a random variable. Shannon (1949) defines the differential entropy H of a random variable X and probability mass function $p(x)$ as

$$H(X) = -\int p(x) \log p(x) dx$$

Since the true probability is not known, it is not possible to calculate $H(X)$ directly.

Entropy estimation has many important applications. For example, it can be used to estimate the mutual information of two random variables and provide insights about their relationship. Furthermore, information entropy has other applications in encoding data, data compression, clustering, and a criterion for feature-splitting in decision trees.

Entropy estimation is difficult because it requires estimating the non-smooth function $f(x) = -x \log(x)$, that is not differentiable at $x = 0$. One approach is to use the naive plugin estimator from the empirical distribution and get

$$\hat{H}(X) = -\sum_{i=1}^{n} \hat{p}(x_i) \text{log} \hat{p}(x_i)$$

where $\hat{p}(x_i) = \dfrac{h_i}{n}$ is the MLE of each probability $p(x_i)$ and $h_i = \displaystyle\sum_{k=1}^{n} \mathbb{I}(X_k = i)$ is the histogram over the outcomes.

However, Basharin (1959) and Harris (1975) have shown that the naive plugin estimator always underestimates the true entropy. Another result from Paninski (2003) proves that there exists no unbiased estimator for entropy. Furthermore, many of the existing estimators suffer from the curse of dimensionality and converge slowly at the rate of $O(T^{-\gamma/d})$ where T is the number of samples and γ is a positive rate parameter.

In this paper, we will summarize the results of some well-known entropy estimators as well as a weighted ensemble method that, under the right conditions, can remove the dependency on the dimension and ensure a convergence rate of $O(T^{-1})$. At the end, we will also discuss the results concerning the minimax rate of entropy estimators.

2 Entropy Estimation Overview

2.1 Plug-in Estimators

Beirlant et al. [1] gives an overview of several popular methods for the nonparametric estimation of entropy. Besides the MLE plug-in estimator, there are other plug-in estimators that utilize techniques such as kernel density estimation, splitting data, or cross-validation.

Consider $X_1, ...X_n$ and an estimator of the form $H_n = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \log f_n(X_i)$ where f_n is a **kernel density estimator**. Joe (1989) shows that the MSE converges at the rate $O(n^{-1}) + O(n^{-2}h^{8-d}) + O(n^{-2}h^{-d}) + O(n^{-1}h^{8-d})+)(n^{-2}h^{4-2d}) + O(h^8)$ where n is the sample size, h is the binwidth, and d is the dimension. The analysis shows that the sample size that is needed for good estimators increases rapidly with the dimension of the multivariate density.

Another plug-in estimator is the **splitting-data estimator**. Start by decomposing the sample into two sets: $X_1, ..., X_i$ and $X_{i+1*}, ..., X_{n*}$. Using the first set of data, construct a density estimate f_i and then, using f_i and the second set, construct $H_n = -\frac{1}{n-(i+1)} \sum_{j=i+1}^{n} \mathbb{I}_{X_{j*}} \log f_j(X_j*)$. For f_j being the histogram density estimate, kernel density estimate, and any L_1-consistent density estimator, Gyorfi and van der Meulen (1987, 1989, 1990) show that the estimator H_n converges almost surely to $H(f)$ under some mild tail and smoothness condition.

The final plug-in estimator is based on **cross-validation**. Let $f_{n,i}$ denotes the kernel density estimator based on $X_1, ..., X_n$ leaving X_i out, then the corresponding density estimator is $H_n = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i} \log f_{n,i}(X_i)$.

2.2 Estimators using Nearest Neighbors

Another interesting technique to estimate entropy is by using the nearest neighbor distances. Let $\rho_{n,i}$ be the nearest neighbor distance of X_i and the other X_j : $\rho_{n,i} = \min_{j \neq i, j \leq n} ||X_i - X_j||$. The nearest neighbor estimate is $H_n = \frac{1}{n} \sum_{i=1}^{n} \ln(n\rho_{n,i}) + \ln(2) + C_E$ where the Euler constant $C_E = -\int_0^\infty e^{-t} \ln(t) dt$.

3 Results

The estimators mentioned above can converge at a very slow rate as the dimension grows large. Sricharan et al. [2] propose a weighted density functionals estimator that takes the weighted affine combination of an ensemble of slow functional estimators to estimate entropy. The ensemble, under the right conditions, can converge at a rate that is independent of the dimension.

Result 1: MSE rate of the weighted ensemble of estimators

Let T be the total sample size, γ be a positive rate parameter, and d be the dimension. Sricharan et al. [2] show that an ensemble of slow estimators with the MSE decay rate of order $O(T^{-\gamma/d})$ can converge at a much faster dimension-invariant rate of $O(1/T)$.

Result 2: Entropy estimation using a weighted ensemble of truncated uniform kernel density estimators

They also apply this weighted estimator, with an ensemble of truncated uniform kernel density estimators, to the problem of Shannon entropy estimation and show that it performs better than many well-known plugin estimators, especially in higher dimensions. Specifically, the weighted ensemble estimator can estimate entropy at $O(T^{-1})$ convergence rate.

Result 3: Entropy estimation using a weighted ensemble of k-nearest neighbors estimators

Gao et al [3] show that k-NN based multivariate entropy estimators [4] achieve parametric MSE rate when the dimension of each of the random variables is less than 3. However, for larger dimensions, Singh and Poczos [5] derive the convergence rates for fixed k-NN entropy estimators. It is shown that the bound on the bias of the estimator is given by $O(\frac{k}{T}^{\beta/D})$, where k is the numbers of neighbors evaluated, β is a measure of the Hölder smoothness of the sampling density, and $D \leq d$ is the intrinsic dimension of the support of the distribution. Since k and β can be set to a constant values and $D < d$, the bias can be further bounded by $O(T^{-d})$. The variance of the estimator is also shown to be bounded by $O(T^{-1})$, so we now have the desirable properties to calculate a weighted ensemble of these k-NN based entropy estimators with parametric MSE rate of $O(T^{-1})$.

Result 4: Minimax rate of entropy estimator

Finally, we look at the minimax rate of estimating integral functions in general. Birge and Massart [6] extend a minimax rate result to estimating functions of the general form $\int \phi(f(x), f'(x), ..., f^{(k)}(x), x)dx$ where f has some density smoothness s. Given the critical index of smoothness $s_c = 2k + d/4$, if $s > s_c$ then an estimator can be constructed to obtain a convergence rate of $O(\frac{1}{\sqrt{n}})$. Otherwise, it is not possible to estimate at a better rate than $O(\frac{1}{n^\gamma})$ where $\gamma = 4s/(4s + d)$ for $d > 1$.

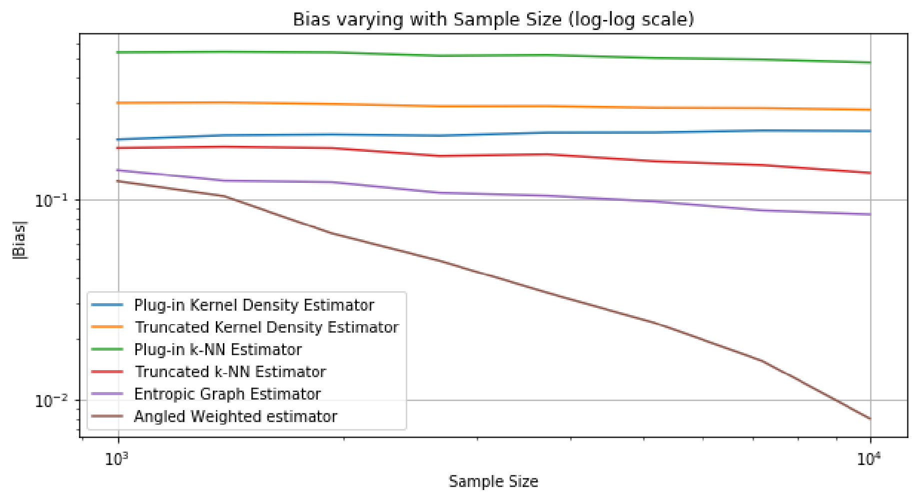

Figure 1: MSE rates for entropy estimators as a function of sample size

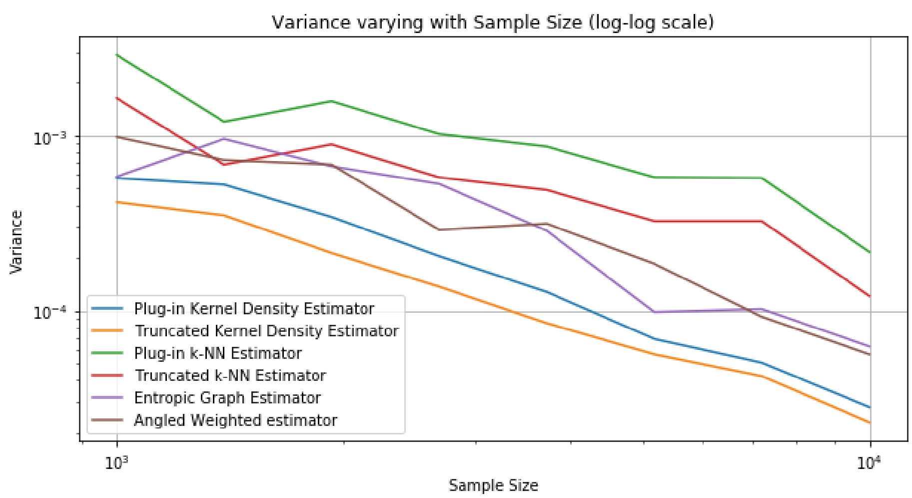

Figure 2: Magnitude of bias for entropy estimators as a function of sample size

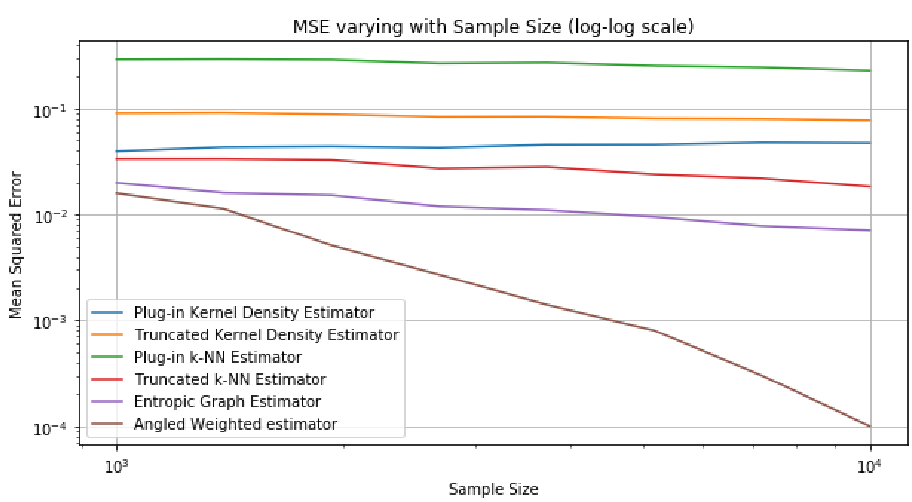

Figure 3: Variance for entropy estimators as a function of sample size

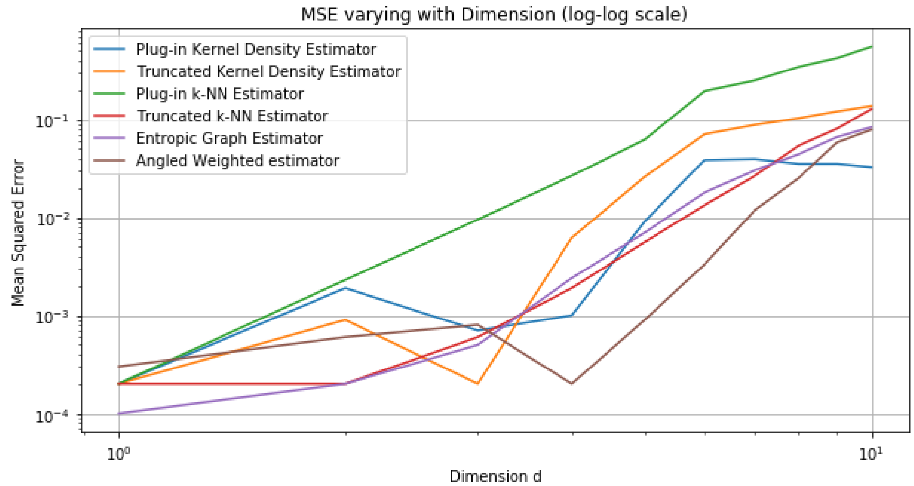

Figure 4: MSE rates for entropy estimators as a function of sample dimension

4 Experiments

Sricharan et al. [2] observe that the weighted ensemble of truncated kernel density estimators achieve mean-squared error rates better than other popular entropy estimators, including the kernel density estimator [7], truncated kernel density estimator, k-NN based Kozachenko-Leonenko estimator [8], truncated KL estimator, and the entropic graph estimator [9].

We used the aforementioned estimators to approximate the Shannon entropy for testing the probability distribution of a random sample and show the MSE rates for each estimator as a function of sample size T in Figure 3, for fixed dimension d = 6. We can see that the weighted ensemble estimator achieves better error rates than those of other estimators. The ensemble converges faster for the bias, whereas for the variance, it converges at a rate similar to those of the other estimators. The experimental results follow our derived results for the bias and variance bounds of the truncated kernel density estimator, weighted estimator, and k-NN based KL estimator—we get similar $O(T^{-1})$ variance bounds for the three estimators, but the weighted estimator achieves a lower bias bound which is independent on the dimension of the data.

We also tried to reproduce the MSE rate results from [2] by varying the sample data dimension d with fixed sample size of 3000 in Figure 4, but were not able to achieve the noticeably superior performances reported in the paper. Specifically, the MSE rates for the weighted entropy estimator deteriorates when $d > 8$. We still observed the expected MSE rate behavior for $d \leq 8$ in that the weighted estimator performs worse than other entropy estimators in lower dimensions, but does better in higher dimensions—this is because the MSE rates for the other estimators depend on d, and thus suffer from curse of dimensionality as d increases.

5 Outline of Proofs

5.1 MSE rate of the weighted ensemble of estimators

We define our ensemble estimator as $\hat{\mathbf{E}}_w = \sum_{l \in \bar{l}} w(l) \hat{\mathbf{E}}_l$, where $\bar{l}$ denotes the set of parameters and $\hat{\mathbf{E}}_l$ are the L parameterized, unbiased estimators. We also add the weights constraint that $\sum_{l \in \bar{l}} w(l) = 1$, which guarantees that the weighted estimator will be asymptotically unbiased. For each $\hat{\mathbf{E}}_l$, we assume that bias and variance is given by:

$$\mathbb{B}(\hat{\mathbf{E}}_l) = \sum_{i \in [I]} c_i \psi_i(l) T^{-i/2d} + O\left(\frac{1}{\sqrt{T}}\right)$$

$$\mathbb{V}(\hat{\mathbf{E}}_t) = c_v \left(\frac{1}{T}\right) + o\left(\frac{1}{T}\right)$$

where c_i and c_v are constants, I is a finite index set with cardinality $I < L$, and $\psi_i(l)$ are basis functions which only depend on l.

Proof: With above bias and variance conditions, a weight vector w_0 can be found by solving a convex optimization problem such that $\mathbb{E}[(\hat{\mathbf{E}}_{w_o} - E)^2] = O(1/T)$. The optimization problem is defined as:

$$\begin{aligned} \min_{w} \quad & ||w||_2 \\ \text{subject to} \sum_{l \in \bar{l}} w(l) \quad & = \quad 1, \\ \gamma_w(i) \quad & = \quad \sum_{l \in \bar{l}} w_l \psi_i(l) = 0, i \in [I], \end{aligned}$$

From this setting, the bias and variance of the weighted estimator can be calculated as:

$$\begin{aligned} \mathbb{B}(\hat{\mathbf{E}}_w) &= \sum_{i \in I} c_i \gamma_w(i) T^{-i/2d} + O\left(\frac{||w||_1}{\sqrt{T}}\right) \\ &= \sum_{i \in I} c_i \gamma_w(i) T^{-i/2d} + O\left(\frac{\sqrt{L}||w||_2}{\sqrt{T}}\right) \end{aligned}$$

Denote the covariance matrix $\{\hat{\mathbf{E}}_l; l \in \bar{l}\}$ by Σ_L. Let $\bar{\Sigma}_L = \Sigma_L T$,

$$\mathbb{V}(\hat{\mathbf{E}}_w) = \mathbb{V}\left(\sum_{l \in \bar{l}} w_l \hat{\mathbf{E}}_l\right) = w^T \Sigma_L w = \frac{w^T \bar{\Sigma}_L w}{T} \leq \frac{L||w||_2^2}{T}$$

We can also rewrite the earlier convex optimization constraints in matrix form as $\min_w ||w||_2^2$ subject to $A_0 w = b$, where b is vector of zeros with a leading one and A_0 is the basis projection matrix—where first row is a vector of ones and the i^{th} row $(A_0)_i = [\psi_i(l_1), ..., \psi_i(l_L)]$. Also, let A_1 be the A_0 matrix without the first row of ones, then solving for the minimum square norm $\eta_L(d) := ||w_0||_2^2$ is given by $\eta_L(d) = \dfrac{\det(A_1 A_1')}{\det(A_0 A_0')}$ where $\eta_L(d)$ is independent of T given fixed number of estimators L and fixed dimension d, i.e. $\eta_L(d) = \Theta(1)$. The bias and variance of the weighted estimator is therefore:

$$\mathbb{B}[\hat{\mathbf{E}}_{w_0}] = O\left(\sqrt{L\eta_L(d)/T}\right) = O\left(1/\sqrt{T}\right)$$

$$\mathbb{V}[\hat{\mathbf{E}}_{w_0}] = O(L\eta_L(d)/T) = O\left(1/T\right)$$

Thus, the overall MSE rate converges in dimension invariant rate of $O\left(1/T\right)$.

5.2 Entropy estimation using a weighted ensemble of truncated uniform kernel density estimators

We can now estimate any general d-dimensional, non-linear density functionals of the form

$$G(f) = \int g(f(x), x) f(x) dx.$$

Let $T = N + M$ and represent the i.i.d observations from f as $\{X_1, ..., X_N, X_{N+1}, ..., X_{N+M}\}$. We have shown an ensemble estimator that will converge much faster than their slower estimator components, if they match the bias and variance conditions. We present one such plugin estimator called the truncated uniform kernel density estimator:

$$\hat{f}_k(X) = \frac{\sum_{i=1}^{M} \mathbb{I}_{\{X_i \in S_k(X)\}}}{M V_k(X)}$$

where $k \leq M$ is a real number, $S_k(X)$ is the truncated kernel region for each X in finite support $[a, b]^d$, and V_k is the volume of the truncated uniform kernel under $S_k(X)$.

Given that we form the estimate $\hat{f}_k$ at N points using M observations, the plug-in estimator is given by:

$$\hat{G}_k = \frac{1}{N}\sum_{i=1}^{N} g(\hat{f}_k(X_i), X_i)$$

where $\hat{G}_k$ is identical to the standard kernel density estimator G'_k except for the volume—which in KDE is set to the untruncated value $V_k(X) = k/M$. It can be shown that the biases of the plug-in estimators $\hat{G}_k, G'_k$ are given by

$$\mathbb{B}(\hat{G}_k) = \sum_{i=1}^{d} c_{1,i}\left(\frac{k}{M}\right)^{i/d} + \frac{c_2}{k} + o\left(\frac{1}{k}\right) + \frac{k}{M}$$

$$\mathbb{B}(G'_k) = c_1\left(\frac{k}{M}\right)^{1/d} + \frac{c_2}{k} + o\left(\frac{1}{k} + \frac{k}{M}\right)$$

The variances of the plug-in estimators $\hat{G}_k, G'_k$ are identical up to leading terms and are given by

$$\mathbb{V}(\hat{G}_k) = \mathbb{V}(G'_k) = c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right)$$

where c_4 and c_5 are constants that depend on g and f. We also need the extra assumptions that $k \to \infty$ and $k/M \to 0$ to ensure that estimators are unbiased. We can see that as $N \to \infty$ and $M \to \infty$, variance converges to 0.

To minimize the bias in both plugin estimators, the optimal choice of $k = \Theta\left(M^{\frac{1}{1+d}}\right)$, which gives us a bias of $\Theta\left(M^{\frac{-1}{1+d}}\right)$. Since $N, M \le T$, both KDE and truncated KDE have the assumed bias and variance rate necessary for the weighted estimator to achieve parametric MSE rate of convergence.

5.3 Minimax rate of entropy estimator

Without loss of generality, the analysis is restricted to functions supported by $[0,1]$. Define $K = (K'_0, ..., K'_m, K_0, ..., K_m)$ for $K'_i \le K_i$ and $0 \le i \le m$. Construct

$$\mathcal{C}^i_m(K) = \{f \in \mathcal{C}^i_m : K'_i \le f^{(i)}(x) \le K_i, \forall x \in [0,1], 0 \le i \le m\}$$

and also for $0 < \nu \le 1$ and $t, A \in \mathbb{R}_+$, construct

$$\mathcal{L}^j_{m+\nu}(K, A) = \{f \in \mathcal{C}^j_m(K) : |f^{(m)}(y) - f^{(m)}(x)| \le A|y-x|^\nu, \forall x, y \in [0,1]\},$$
$$\tilde{\mathcal{L}}_{m+\nu}(t, A) = \{f \in \mathcal{L}^\infty_{m+\nu}(tU, A) : \textstyle\int_0^1 f(x)dx = 0\}$$

with $U =$ **-1**$_{m+1}$, $\mathbf{1}_{m+1}$ where $\mathbf{1}_{m+1}$ has all components equal to 1 in $\mathcal{R}^{m+1}$.

We also make the following assumption

Assumption $\mathbb{A}(f, \mu)$: There exist disjoint sets $A_1, ..., A_p$ and functions g_i satisfying the following relations for $1 \le i \le p$:

i. $||g_i||_\infty \le 1$;
ii. $||\mathbb{I}_{A_i} g_i||_1 = 0$;
iii. $\int g_i(x) f(x) d\mu(x) = 0$;
iv. $\int g_i^2(x) f(x) d\mu(x) = a_i > 0$

Theorem 1: Let $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_p\}$ where $\lambda_i \in \Lambda = \{-1, 1\}^p$. Define $\bar{Q}_n = 2^{-p}\sum_{\lambda \in \Lambda} Q^n_\lambda$ and assuming $\mathbb{A}(f, \mu)$ is satisfied. Also let

$$\alpha = \sup_{1 \le i \le p} ||g_i||_\infty, \qquad s = (n\alpha^2)\sup_{1 \le i \le p} P(A_i), \qquad c = (n)\sup_{1 \le i \le p} a_i$$

Then, with $C \le \frac{1}{3}$, we can upper bound the Hellinger distance by

$$h^2(P^n, \bar{Q}_n) \leq C(\alpha, s, c)n^2 \sum_{j=1}^{p} a_j^2$$

The proof is lengthy and is omitted here. We can use this result and the Le Cam method to obtain the lower bound.

Corollary 1: Assume that T is a function defined on some subset Θ of $\mathbb{L}^1(\mu)$, which contains f and some set of densities $g_\lambda, \lambda \in \Lambda$, derived from g_i's which satisfy $\mathbb{A}(f, \mu)$ with parameters α, s, c as defined above. If (i) $C(\alpha, s, c)n^2 \sum_{j=1}^{p} \alpha_j^2 \leq \gamma < 1$ and (ii) $\forall \lambda \in \Lambda, T(g_\lambda) - T(f) \geq 2\beta > 0$, then for any estimator $\hat{T}_n$ of T derived from n i.i.d. observations, we have for the joint distribution $\mathbb{P}_g = g \cdot \mu$

$$\sup_{g\in\Theta} \mathbb{P}_g[|T(g) - \hat{T}_n| > \beta] \geq \tfrac{1}{2}[1 - (\gamma(2-\gamma))^{1/2}]$$

Proof: Suppose $T(f) = 0$, define subsets Θ_0 and Θ_1 of Θ as

$$\Theta_0 = \{g \in \Theta : T(g) \leq 0\}, \qquad \Theta_1 = \{g \in \Theta : T(g) \geq 2\beta$$

By the Le Cam method, any test between Θ_0 and Θ_1 will have at least one of its errors as large as $(1/2)[1 - (\gamma(2-\gamma))^{1/2}]$. If we consider the particular test which accepts Θ_0 when $\hat{T}_n \leq \beta$, we get the same result as in Corollary 1.

To prove the lower bound in the multi-dimensional case, we will work on the hypercube $H = [0,1]^d$ and consider functions in $\tilde{L}_{m+\nu}(t, A)$.

Theorem 2: Let f be some density in Θ and $\log f$ be bounded on H. Denote B_t as the set $\{f + l | l \in \tilde{L}_{m+\nu}(t, A)\}$ where m, ν, A are fixed constants and assume that $B_t \subset \Theta$ for some small t and that when $f + l \in B_t$, we can expand

$$T(f+l) = T(f) + \int_H T'(x)l(x)dx + \tfrac{1}{2}\int_H T''(x)l^2(x)dx + ||l||_2^2 o(1)$$

where $o(1)$ is a function of t and $\inf_{x\in H} T''(x) > 0$. Then

$$\lim_{t\to 0}\inf \lim_{t\to 0}\inf \inf_{\hat{T}_n} \sup_{g\in B_t} \mathbb{P}_g[|\hat{T}_n - T(g)| > \epsilon n^{-\gamma}] > 0$$

for some $\epsilon > 0$ and $\gamma = 4s/(4s+d)$.

Proof: Divide the hypercube $[0,1]^d$ into $p = \prod_{i=1}^{d} p_i$ hyperrectangles with side lengths p_i^{-1} chosen in such a way that, for some K chosen later, $K \leq A_i p_i^{-(m_i+\nu_i)} \leq 2K, 1 \leq i \leq d$. On each hyperrectangles $R_j, 1 \leq j \leq p$, we can build a perturbation l_j such that for a fixed constant c,

$$\int_{R_j} l_j(x)dx = 0, \qquad \int_{R_j} T'(x)l_j(x)dx = 0, \qquad \int_{R_j} l_j^2(x)dx \geq \tfrac{c^2}{p}$$

$||D_i^{m_i} l_j||_\infty \leq p_i^m$ for $1 \leq i \leq d$. For $\lambda \in \Lambda\{-1; 1\}^p$, setting

$$l_\lambda(x) = K \sum_{j=1}^{p} \lambda_j l_j(x) \mathbb{I}_{R_j}(x)$$

we see that for p large enough, $f + l_\lambda$ belongs to Θ for all λ and that

$$T(f + l_\lambda) \geq T(f) + C_1 K^2$$

We can apply Corollary 1 with $\sum_{j=1}^{p} \alpha_j^2 \leq C_2 K^4/p$. Since p is, by definition of K, of order $K^{-d/s}$, choosing $K = C_3 n^{-2/(4+d/s)}$ gives us the desired result.

For the upper bound, we can construct an estimator $\hat{f}$ of f based on r i.i.d. variables of density f such that for $r \geq r_0$ not depending on f, the following holds:

i. $\hat{f} \in \mathcal{C}_{2k}^0(K^\epsilon)$;

ii. for $2 \geq q < \infty$ and $0 \geq i \geq k$,

$$\mathcal{C}_f(||\hat{f}^{(i)} - f^{(i)}||_q^q \leq C_i'(q) r^{-q/6},$$

for some constants $C'_i(q)$ that are independent of f.

Now we can define the estimator $\hat{T}_n$ of $T(f)$ as follows:

$$\hat{T}_n = T(\hat{f}) - \sum_{i=0}^{k} < \phi'_i(\hat{f}, \hat{f}^{(i)} > + \frac{1}{2} \sum_{i,j=0}^{k} < \phi''_{i,j}(\hat{f}), \hat{f}^{(i)} \hat{f}^{(j)} > + \overline{L(\hat{f})} + \frac{1}{2} \sum_{i,j=0}^{k} Q_{i,j}(\phi''_{i,j}(\hat{f}))$$

where $\phi_1, ..., \phi_q$ is an orthonormal system in $\mathbb{L}^2([0,1], dx)$ which has the following properties:

i. $\phi_1 = 1$ and $\phi_j(x) = 0$ for $j \geq 2, x \notin [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$;

ii. The linear space $\mathbb{V}$ generated by ϕ'_j is stable by differentiation.

Given this estimator, we can use the Taylor expansion and bounding the remainder to show that $\mathbb{E}_f[|\hat{T}_n - T(f)|^2] \leq Cn^{-1}$.

6 Conclusion

In this paper, we have explored the theoretical foundations and performance in practice of an optimally weighted ensemble of slowly-converging entropy estimators. These estimators which have deteriorating MSE rates as the dimension d increases can be used in a weighted ensemble that can achieve the minimax $O(T^{-1})$ rate of convergence. The weights can be determined by solving a convex optimization problem which does not require any training data. We also have shown that the popular kernel density estimator, its bias-corrected version, and the k-NN based Kozachenko-Leonenko estimator fit the conditions of having slow MSE rates of order $O(T^{-\gamma/d})$ and can therefore be used to derive the weighted estimator.

Additionally, we ran some experiments to showcase the weighted ensemble estimator's superior performance in approximating Shannon entropy compared to other widely-used entropy estimators and related the observed results to our theoretical findings. Although we did not observe the same MSE rate when varying over the dimension as [2] reported, we believe that the our experiment performed poorly when the dimension $d > 8$ because the sample size of 3000 may not have been enough to estimate the Shannon entropy, and increasing T should result in our estimators following the same trend as [2] observed. Assuming that the weighted estimators can achieve the $O(T^{-1})$ convergence rate as expected, the ensemble method poses as a promising technique to perform effective entropy estimation.

In closing, entropy estimation is still an open area of research. One possible extension of the weighted estimator algorithm is to use an L_1 norm instead of L_2 to introduce sparsity. This approach may speed up the estimation process since there would be fewer estimators. Furthermore, we can apply the weighted ensemble to estimate other functionals of probability density that has a general form of $\int \phi(f(x), f'(x), ..., f^{(k)}(x), x)dx$ such as divergence, mutual information, and intrinsic dimension.

References

[1] J Beirlant, E J Dudewicz, L Gyorfi, and E C Van Der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[2] K Sricharan, D Wei, and AO Hero. Ensemble estimators for multivariate entropy estimation. *IEEE transactions on information theory / Professional Technical Group on Information Theory*, 59:4374–4388, 7 2013.

[3] W Gao, S Oh, and P Viswanath. Demystifying fixed k-nearest neighbor information estimators. *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1267–1271, 2017.

[4] S Singh and B Poczos. Analysis of k-nearest neighbor distances with application to entropy estimation. 03 2016.

[5] S Singh and B Poczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. 2016.

[6] L Birge and P Massart. Estimation of integral functionals of a density. *Ann. Statist.*, 23(1):11–29, 02 1995.

[7] Berwin A. Turlach. Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*.

[8] K. Fukunaga and L. Hostetler. Optimization of k nearest neighbor density estimates. *IEEE Trans. Inf. Theor.*, 19(3):320–326, May 1973.

[9] K. Sricharan, R. Raich, and A. O. Hero, III. Empirical estimation of entropy functionals with confidence. *ArXiv e-prints*, December 2010.