ENTROPY ESTIMATION USING THE PRINCIPLE OF MAXIMUM ENTROPY

Behrouz Behmardib, Raviv Raich1, and Alfred O. Hero III2

1 School of EECS, Oregon State University, Corvallis, OR 97331
2 Department of EECS, University of Michigan, Ann Arbor, MI 48109
{behmardb, raich} @eecs.oregonstate.edu, hero@umich.edu

ABSTRACT
In this paper, we present a novel entropy estimator for a given set of samples drawn from an unknown probability density function (PDF). Counter to other entropy estimators, the estimator presented here is parametric. The proposed estimator uses the maximum entropy principle to offer an m-term approximation to the underlying distribution and does not rely on local density estimation. The accuracy of the proposed algorithm is analyzed and it is shown that the estimation error is \( \leq O(\sqrt{n}) \). In addition to the analytic results, a numerical evaluation of the estimator on synthetic data as well as on experimental sensor network data is provided. We demonstrate a significant improvement in accuracy relative to other methods.

Index Terms— Maximum entropy, Entropy estimation, m-term approximation

1. INTRODUCTION
Information theory quantities such as entropy and mutual information are widely used in data analysis, signal processing, and machine learning. When an underlying model for data is unavailable, sample-based entropy estimation is required. Entropy estimation has been applied in anomaly detection, image segmentation, estimation of manifold dimension and feature selection (e.g., see [1]). In this paper, we consider estimation of the entropy of a continuous random variable characterized by a PDF. In the discrete case, raw counts are used to estimate the probability for each discrete value and consequently, entropy is estimated using the plug-in method. In the continuous case, two main approaches exist. In the first approach, the PDF is approximated and then the result of the approximation is plugged into the entropy formula (e.g., kernel density, histogram). In the second approach, the entropy is estimated directly from samples (e.g., sample spacing, nearest neighbors, and entropic spanning graph, see [2] for a review).

The main contribution of this paper is the development of a new entropy estimator based on the principle of maximum entropy and greedy m-term approximation. We also analyze the estimation error, specifically an in probability error bound in terms of the problem parameters (e.g., number of samples, number of the approximation terms). The error of the proposed estimator is \( O(\sqrt{n}/n) \); only a factor of \( \sqrt{n} \) away from the classical statistical parameter estimation error \( O(1/n) \). Using numerical examples, we demonstrate that our algorithm is competitive compared with the other well-known algorithms.

This work was partially supported by the Air Force Office of Scientific Research grant number FA9550-09-1-0471.

2. PROBLEM FORMULATION
We consider the estimation of the entropy of random variable \( X \) from \( n \) i.i.d. samples of it. Let \( X \) be a random variable with a PDF \( p(x) \). The entropy of \( X \) is given by

\[
H(p) = -\int p(x) \log p(x) \, dx.
\]

We are interested in an entropy estimator \( \hat{H} : \mathcal{X}^n \rightarrow \mathbb{R} \) of \( H(p) \), which takes \( x_1, x_2, \ldots, x_n \in \mathcal{X} \) as the input. We seek a consistent estimator in the following sense:

\[
\lim_{n \rightarrow \infty} \mathbb{E}_{x_1, \ldots, x_n} \hat{H}_n(x_1, \ldots, x_n) \longrightarrow H(p) \quad \text{in probability.}
\]

We are also interested in quantification of the estimation error \( H(p) - \hat{H}(p) \).

3. SOLUTION FRAMEWORK
To estimate the entropy given in (1), two approximations are typically considered. The first involves replacing the expectation with a sample average. The second involves the more challenging task of estimating \( p(x) \) or \( \log p(x) \). To address the second approximation, we consider a maximum entropy approach to model \( p(x) \).

3.1. Maximum entropy framework for entropy estimation
Assume that one has access to the expected value of \( m \) different features \( \{\phi_j(x)\}_{j=1}^m \) (e.g., mean \( E[x] \) and second-order moment \( E[x^2] \)) w.r.t. to PDF \( p(x) \). Even if \( m \) is large, one cannot identify \( p(x) \) uniquely. Maximum entropy principle allows for finding a unique distribution among all distributions that satisfy a set of constraints:

\[
\max_{p} H(p) \quad \text{s.t.} \quad E_p[\phi_j(x)] = \alpha_j, \quad j = 1, 2, \ldots, m,
\]

where \( H(p) \) is given in (1), \( \phi_j(x) \) is a feature function, and \( \alpha_j \) is the expected value of the \( j \)th feature. The distribution that solves the
where $\lambda \in \mathbb{R}^m$ is the solution to the set of equations $E_{p_\lambda} [\phi_j(x)] = \alpha_j$ for $j = 1, 2, \ldots, m$ and $Z(\lambda) = \log \int \exp \left( \sum_{j=1}^m \lambda_j \phi_j(x) \right) dx$ [3]. Substituting the PDF given by (4) into (1), yields a parametric expression for the entropy:

$$H(p_\lambda) = Z(\lambda) = \sum_{j=1}^m \lambda_j E_{p_\lambda} [\phi_j(x)].$$

The set of PDFs $P = \{p_\lambda : \lambda \in \mathbb{R}^m\}$ provides an approximation space for $p$. The set $P$ is convex [3] and as a results, a unique $p_{\lambda^*} \in P$ can be found, which minimizes the Kullback-Leibler (KL) divergence between the distribution $p$ and its approximation $p_{\lambda^*}$ given by

$$D(p||p_{\lambda^*}) = H(p) - H(p_{\lambda^*}) \leq c/m \quad [4].$$

This approximation capability of the maximum entropy framework is key to our method suggesting the idea of replacing $H(p)$ with $H(p_{\lambda^*})$.

Since only observations $x_1, x_2, \ldots, x_n$ from $p(x)$ are available, one cannot obtain $\lambda^*$ directly based on $p(x)$. Instead, $\hat{\lambda}$ is obtained by maximizing the likelihood or equivalently by minimizing the negative log-likelihood $Z(\lambda) = \sum_{j=1}^m \lambda_j E_{\hat{p}} [\phi_j(x)]$, where $\hat{p}$ is the empirical distribution for which $E_{\hat{p}} [f(x)] = \frac{1}{n} \sum_{i=1}^n f(x_i)$. The entropy $H(p_{\lambda^*})$ is given by

$$H(p_{\lambda^*}) = \min_{\lambda} Z(\lambda) = \sum_{j=1}^m \lambda_j E_{\hat{p}} [\phi_j(x)].$$

The sample-based entropy estimated in (7) provides an estimate to (6). By concentration of measure, i.e., the property that $E_{\hat{p}} [f(x)] \to E_{p} [f(x)]$, one can show that (7) converges to (6) in probability. Motivated by the approximation and estimation capabilities of the framework, we proceed with the description of two specific estimators and their properties.

### 3.2. Proposed estimators

There are two key issues which have to be addressed in finding an optimum estimator for the entropy using the framework of maximum entropy. The first issue is to find the optimum $\lambda$ in (7) which can be done by a variety of convex optimization tools. Specifically for this model, iterative scaling is a common approach [5]. The second issue is to find the best set of $\phi$’s which provides an accurate approximation for the true entropy. For that end, we define a collection of feature functions $\phi$ given by $\Phi = \{\phi_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^d$. Suppose $\phi_{\theta^1}, \ldots, \phi_{\theta^m}$ are the features used to approximate $p(x)$. Thus, the entropy estimator is

$$\hat{H}^{(m)}(\theta_1, \ldots, \theta_m) = \min_{\lambda} Z(\lambda; \theta) = \sum_{i=1}^m \lambda_i E_{\hat{p}} [\phi_{\theta_i}(x)].$$

While the solution to $\lambda$ is straightforward following the maximum entropy approach, the choice of $\theta$ is not trivial. Two estimators are proposed to address the selection of $\theta$ in (8) and analysis of the error is provided.

#### 3.2.1. Brute-force $m$-term entropy estimator

We propose the following estimator

$$\hat{H}^{(m)}_1 = \min_{\theta_1, \ldots, \theta_m} Z(\lambda; \theta) - \sum_{i=1}^m \lambda_i E_{\hat{p}} [\phi_{\theta_i}(x)].$$

The solution to (9) presents a strategy for finding the $\phi_i$ in (7). The joint minimization of $\theta_1, \ldots, \theta_m$ presents a computational challenge. However, the estimator performance can allow us to understand the limitations of the approach.

**Theorem 3.1** Let $\hat{H}^{(m)}_1 = \hat{H}^{(m)} - C/2m$. The estimation error associated with $\hat{H}^{(m)}_1$ satisfies:

$$|\hat{H}^{(m)}_1 - H(p)| \leq C \sqrt{\frac{2m}{n \log \frac{2m}{\delta}}},$$

with probability at least $1 - \delta$, where $C = \frac{1}{2} \left[ \log p - \log p_{\lambda^*} \right] \leq M$, and $\|\lambda\|_1 \leq L$.

Theorem 3.1 decomposes the error of estimating the entropy into two parts: approximation error and estimation error (analogous to the familiar bias and variance decomposition in classical statistics). The first term on the RHS is corresponding to approximation error. Increasing the number of terms $m$ provides a rich basis for the space that includes the target function $\log p(x)$ and hence reduces the error. Simultaneously, the estimation error is increased. The second term is the estimation error which decreases as the number of samples $n$ increases. Constant $C$ depends on $\|f\|_{\infty}$ where $f(x) = \log p(x)$. Common in approximation theory, the approximate function $f(x)$ is assumed to be bounded $\|f\|_{\infty} \leq M$. The details of the derivation of parameter $C$ are given in [4]. Due to space limitation the details of the proof are provided in [6]. However, we proceed with some intuition. Consider the decomposing the error as:

$$\hat{H}^{(m)}_1 - H(p) = \min_{\theta_1, \ldots, \theta_m} D(p||p_{\lambda^*}) + \sum_{i=1}^m \lambda_i E_{\hat{p}} - \hat{p}, [\phi_{\theta_i}].$$

Barron et. al. has shown in [4] that approximation error can be bounded as $D(p||p_{\lambda^*}) \leq C/m$. Hoefding’s inequality provides a bound on the difference between empirical mean and true mean of a function of i.i.d. bounded random variable [7]. By applying Hoefding inequality to the estimation error we can bound the estimation error by $C \sqrt{\frac{m}{n \log \frac{2m}{\delta}}}$. To find the rate of convergence based on the number of the samples $n$, we present the following corollary:

**Corollary 3.2** Let the number of features used to approximate $p(x)$ be $m = \sqrt{n}$, then with probability $1 - \delta$ the estimation error is bounded by

$$|\hat{H}^{(m)}_1 - H(p)| \leq C_1 \sqrt{\frac{\log n}{n}} + o\left(\sqrt{\frac{\log n}{n}}\right),$$

where $C_1 = CML \sqrt{\frac{1}{\delta} \log \frac{2}{\delta}}$.
This corollary suggests that the overall error is $O(\sqrt{\log n/n})$; only a factor of $\sqrt{\log n}$ away from the statistical estimation error $O(\sqrt{1/n})$. While computationally demanding, the performance of the proposed estimator illustrates the merit in the maximum entropy framework. Note that this bound can be improved which means the proposed bound is not necessary the tightest bound.

### 3.2.2. Greedy m-term entropy estimator

Greedy approaches for approximating functions with $m$-terms from a given dictionary $D$ were shown to be effective [8]. Greedy $m$-term approximations offer a computationally efficient alternative to joint optimization of $m$-term approximations. We consider the greedy approach for the following entropy estimator due to its computational efficiency. We would like to arrive to the $m$-term approximation of $\log p(x)$, of the form $g_m(x) = \sum_{j=1}^{m} \lambda_j \phi_{y_j}(x)$ by adding one term at a time. Start by initializing $g_0(x) = 0$. The $l$th iteration considers constructing $g_l(x)$ based on $g_{l-1}(x)$ through

$$g_l(x) = (1 - \frac{1}{l})g_{l-1}(x) + \frac{1}{l} \beta \phi_{y_l}(x),$$

where $\beta$ and $\theta$ are obtained by

$$\min_{\beta, \theta} Z(g_l(x)) = E_p[g_l(x)].$$

The minimization in (14) is convex w.r.t. $\beta$ when $\theta$ is held fixed leaving the main difficulty to optimization w.r.t only a single variable $\theta$. After $m$ iterations, we obtain $g_m(x)$ of the form $g_m(x) = \sum_{j=1}^{m} \lambda_j \phi_{y_j}(x)$. Substituting the values of $\{\lambda_j, \phi_{y_j}\}_{j=1}^{m}$ into

$$\hat{H}_2^{(m)} = Z(\lambda) - \sum_{j=1}^{m} \lambda_j E_{p}[\phi_{y_j}],$$

yields the proposed entropy estimate. Despite the potential sub-optimality of the greedy approach, the method provides consistent entropy estimates. Its accuracy is examined in the following theorem.

**Theorem 3.3** For $\hat{H}_2^{(m)}$ defined in (15) and $m = \sqrt{n}$ with probability at least $1 - \delta$,

$$|\hat{H}_2^{(m)} - H(p)| \leq \frac{K_1 \sqrt{d} + 2 \log m}{m} + \frac{K_2}{m} + \frac{K_3}{m^2}$$

where $d = \sqrt{2 \log 2}$, $K_1 = 8LM$, $K_2 = 16L^2M^2$, and $K_3 = 64C_3L^3M^3$, and $C_2 = 6(eL - 1 - LM - L^2M^2/2)/L^3M^3$.

Due to space limitation, we omit the proof for this theorem, which is available in [6]. Similar to Theorem 3.1, this bound decomposes the error into approximation error and estimation error. The first term on the RHS is related to the estimation error while the second and third terms are related to the approximation error. We proceed with a corollary, which expresses the convergence rate of the algorithm in terms of the number of samples $n$.

**Corollary 3.4** If we select the number of terms $m$ as $m = \sqrt{n}$ in $\hat{H}_2^{(m)}$ from (15), the estimation error of $\hat{H}_2^{(m)}$ is bounded with probability $1 - \delta$ by

$$|\hat{H}_2^{(m)} - H(p)| \leq C_3 \sqrt{\frac{\log n}{n}} + o\left(\frac{\log n}{n}\right),$$

where $C_3 = K_1 + K_2 + K_3$.

While the greedy method is typically expected to present performance inferior to that of the brute-force estimator, its asymptotic error is of the same order. From a computational point of view, the greedy approach is significantly faster than the brute force method. We proceed with the computationally efficient greedy $m$-term estimator. In the next section, the performance of the estimator is numerically evaluated and compared to alternatives.

### 4. SIMULATIONS

In this part we compare the performance of well known entropy estimation approaches with the greedy $m$-term estimator defined in Section 3.2.2 on data drawn from three univariate continuous distributions as well as on experimental sensor network data. The estimators considered in this comparison are: (i) *Histogram*: the plug-in estimator for the histogram density estimation using a constant bins width chosen according to [9]. (ii) *KDE*: the kernel density estimator with the optimal bandwidth chosen according to [10].

(iii) *Sample spacing*: the classical sample spacing approach with $m = 5$. (iv) *Nearest neighbors*: the nearest neighbor estimator with $k = 5$. (v) *Greedy $m$-term*: the proposed approach with a dictionary of 1500 features $\phi$ including polynomials $x^2$ and trigonometric basis $\sin(2\pi x), \cos(2\pi x)$ for $i = 1, \ldots, 500$ and optimum number of $m$ varying between 10 to 50.

#### 4.1. Synthetic dataset

We consider three univariate distributions: *truncated normal* with $\mu = 0.5$ and $\sigma = 0.2$, *uniform* between $(0, 1)$, and truncated mixture of five Gaussians with $\mu = [0.3, 0.5, 0.7, 0.8, 0.85]$ and $\sigma = [0.09, 0.01, 0.009, 0.001, 0.0005]$ respectively. For each distribution, samples of size $[100, 200, 500, 1000, 2000]$ were considered and 10 runs of the experiment were conducted. The left column of Fig. 2 depicts the distributions and the right column shows the accuracy of algorithms in terms of mean square error. For the two simple classical example (truncated normal, uniform) all algorithms perform very closely. However, in the mixture of Gaussians example $m$-term estimator significantly outperforms other algorithms. Note that there is no Gaussian basis in the dictionary $D$. The approximation of five mixture of Gaussians using the $m$-term approximation was performed and the result is depicted in Fig. 3. This example illustrates the approximation of the true density. The figure is in log scale since $\log p(x)$ is approximated by a linear combination of the features $\phi_j$’s.

#### 4.2. Anomaly detection in sensor network

We considered the use of the greedy $m$-term estimator for anomaly detection. An experiment was set up on a Mica2 platform, which consists of 14 sensor nodes randomly deployed inside and outside a lab room. Wireless sensors communicate with each other by broadcasting and the received signal strength (RSS), defined as the voltage measured by a receiver’s received signal strength indicator circuit (RSSI), was recorded for each pair of transmitting and receiving nodes. There were $14 \times 13 = 182$ pairs of RSSI measurements over a 30 minute period, and each sample was acquired every 0.5 sec. During the measuring period, students walked into and out of lab at random times, which caused anomaly patterns in the RSSI measurements. Finally, a web camera was employed to record activity for ground truth. The mission of this experiment is to use the 182 RSS sequences to detect any intruders (anomalies). Fig. 4 shows the results of the greedy $m$-term estimator and nearest neighbor. Due to
Fig. 2. Toy examples. Left: PDF $p(x)$. Right: MSE of five different estimators of $H(p)$ vs. sample size $N$.

Fig. 3. Graph of $\log p(x)$ vs. the approximated $\log p(x)$ using $m$-term approximation approach.

space limitation, we omitted the results of other algorithms on this dataset. We observe that the entropy peaks at times of anomaly in a similar fashion for both methods. Though the two methods are based on different frameworks, similar entropy estimates are produced.

5. CONCLUSION

In this paper, we proposed the maximum entropy framework for entropy estimation. The proposed estimators deploy $m$-term approximation to estimate the entropy. In addition to a brute-force estimator, we introduced a low computational complexity greedy $m$-term entropy estimator. Theoretical analysis of the proposed estimators shows that the estimation error is $O(\sqrt{\log n/n})$. As with other entropy estimation methods, the proposed method can be used for a variety of applications. The application of the method to anomaly detection in sensor networks was demonstrated. Our proposed estimator was shown to be competitive with other approaches. Future work may address some of the following: i) the choice of dictionary and ii) extension to unbounded support.

6. REFERENCES