Estimation of the Entropy on the Basis of its Polynomial Representation

Martin Vinck*, Francesco P. Battaglia*, Vladimir B. Balakirsky†, A. J. Han Vinck† and Cyriel Pennartz*
*Center for Neuroscience
University of Amsterdam
Amsterdam, The Netherlands, 1098 XH
martinvinck@gmail.com
†Institute for Experimental Mathematics
University of Essen
Essen, Germany, 45326

***Abstract*—An algorithm for estimating the entropy, which is based on the representation of the entropy function as the sum of two polynomial terms, called the polynomial approximation function and the remainder, is proposed. We construct an accurate and unbiased estimate of the value of the polynomial approximation function and use the known Bayesian approach to estimate the remainder. The combined estimator essentially reduces the bias of the constructed estimate as compared to the known estimators. Simulation results that confirm the claim are presented.**

I. Introduction

We consider the estimation of the entropy of the probability distributions for discrete memoryless sources when a limited amount of observed data is available. This mathematical problem is important for many applications. In particular, it received a great attention in neurophysiology where the performance of networks of neurons is evaluated by information theoretical quantities [1]. Plugging in the empirical frequencies into the entropy function yields a direct estimate of entropy, which is strongly negatively biased by sample size. This negative estimator bias can translate into a strong positive bias in mutual information. This is a serious issue for fields such as neuroscience, where obtaining large amounts of data is virtually impossible, because of technical limitations in maintaining stable monitoring of neural signals for long times, or because the brain remains in the same state only for short periods.

Searching for an entropy estimator with minimum bias or distortion leads to large variance and asymptotical efficiency issues, and there is a general trade-off between the variance and bias of entropy estimators [2]. A particularly promising framework to estimate entropy is the Bayesian one [3], [4]. Wolpert and Wolf [4] developed a Bayesian estimator based on a uniform prior distribution of probabilities. Nemenman et al. [3], [5] pointed out that this uniform prior distribution of probabilities corresponds to a peaked and informative prior on the entropy with an expected value that is relatively close to maximum entropy, leading to large errors when the actual value of the entropy is small. Therefore, the Bayesian Nemenman-Shafee-Bialek (NSB) estimator uses a nearly flat prior distribution over the values of the entropy constructed as a mixture of symmetric Dirichlet distributions [3]. The NSB estimator exhibits rapid convergence to the entropy and has good performance in terms of bias and robustness in comparison to other available estimators [1], [3], although systematic comparisons between available estimators are lacking.

In the present correspondence, we show that further improvements can be made on the NSB estimator in terms of estimator bias. The paper is organized as follows. We use the Taylor power series for the logarithm function to decompose the entropy into two terms, called the polynomial approximation term and the remainder. An accurate and unbiased estimate for the polynomial approximation term, which is of interest by itself, is presented. The remainder is estimated using the known Bayesian approaches. The combined estimator has essentially better performance than the known estimators. This claim is also confirmed by simulation results.

II. Problem of entropy estimation

Suppose that there is a discrete memoryless source which generates one of M values (states or symbols) from the set $\mathcal{R} = \{r_1, \ldots, r_M\}$ specified by the vector of probabilities $\boldsymbol{p} \equiv (p_1, \ldots, p_M)$. The (Shannon) entropy function [6] is defined as $H(\boldsymbol{p}) \equiv -\sum_{m=1}^{M} p_m \ln(p_m)$ (hereafter, we assume that $0 \ln 0 = 0$). The received vector of outcomes of n independent observations is defined as $\boldsymbol{x} \equiv (x_1, \ldots, x_n) \in \mathcal{R}^n$. Let $\boldsymbol{N} \equiv (n_1, \ldots, n_M)$ with the m-th component defined as the number of symbols r_m in the vector $\boldsymbol{x}$. The 'plugin' entropy estimator is defined as $\hat{H}_{\text{plugin}}(\boldsymbol{n}) = -\sum_{m=1}^{M} \frac{n_m}{n} \ln\left(\frac{n_m}{n}\right)$ and it is well-known that it has a strict non–positive bias: As $f(p_m) \equiv -p_m \ln(p_m)$ is a concave function of p_m , the inequality $\mathrm{E}\{-\frac{n_m}{n} \ln(\frac{n_m}{n})\} \leq -\mathrm{E}\{\frac{n_m}{n}\} \ln(\mathrm{E}\{\frac{n_m}{n}\})$ stems from Jensen's inequality.

The problem under consideration is to provide an estimate of $H(\boldsymbol{p})$ based on the observation $\boldsymbol{n}$, $\hat{H}(\boldsymbol{n})$, that has both low bias $|\,\mathrm{E}\{\hat{H}(\boldsymbol{n})\} - H(\boldsymbol{p})|$ and low mean absolute error $\mathrm{E}\{|\hat{H}(\boldsymbol{n}) - H(\boldsymbol{p})|\}$. We want bias and mean absolute error to be small for all vectors of probabilities $\boldsymbol{p}$ (equivalently, for all entropies $H(\boldsymbol{p}) \in [0, \log(M)]$).

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

III. Polynomial representation of the entropy function.

For all $n \geq 2$, we represent the entropy function as the sum

$$H(\boldsymbol{p}) = T(\boldsymbol{p}) + R(\boldsymbol{p})\,. \tag{1}$$

The polynomial approximation $T(\boldsymbol{p})$ is defined as

$$T(\boldsymbol{p}) \equiv \sum_{m=1}^{M} \sum_{k=1}^{n} a_k p_m^k\,, \tag{2}$$

where the coefficients a_k are defined as

$$a_k \equiv \sum_{j=k}^{n-1} \frac{1}{j} \binom{j}{j-k+1} (-1)^{k-1}\,. \tag{3}$$

The remainder function $R(\boldsymbol{p})$ is defined as

$$R(\boldsymbol{p}) \equiv \sum_{m=1}^{M} \sum_{k=0}^{\infty} p_m (1-p_m)^n \frac{(1-p_m)^k}{k+n}\,. \tag{4}$$

Such a polynomial representation follows from the $(n-1)$-th order Taylor expansion of $-\ln(p_m)$ around $p_m = 1$, $-\ln(p_m) \approx \sum_{k=1}^{n-1} \frac{(1-p_m)^k}{k}$. The polynomial approximation of $f(p_m) \equiv -p_m \ln(p_m)$ is then defined as $g(p_m) \equiv \sum_{k=1}^{n-1} p_m \frac{(1-p_m)^k}{k}$. By the binomial theorem, $(1-p_m)^k = \sum_{j=0}^{k} \binom{k}{j} (-p_m)^{k-j}$, and we can represent the approximation function $g(p_m)$ as

$$g(p_m) = \sum_{k=0}^{n-1} \sum_{j=k}^{n-1} (-1)^k \frac{1}{j} \binom{j}{j-k} p_m^{k+1}\,. \tag{5}$$

where we define $\binom{j}{j-k} \equiv 0$ if $j < k$.

The polynomial representation $T(\boldsymbol{p})$ is a meaningful function in itself that shares various properties with the entropy function. (i) Like entropy, $T(\boldsymbol{p})$ is a non-negative function. (ii) Like the entropy function, the function $T(\boldsymbol{p})$ is strictly concave. (iii) Both $H(\boldsymbol{p})$ and $T(\boldsymbol{p})$ attain maximum values when all probabilities are equal, and any change towards equalization of the probabilities increases both of them. (iv) The function $T(\boldsymbol{p})$ is a monotonically increasing function of M when all p_m's are equal to each other. (v) The inequality $T(\boldsymbol{p}) \leq H(\boldsymbol{p})$ holds. (vi) As $n \to \infty$, $T(\boldsymbol{p}) \to H(\boldsymbol{p})$. (vii) The function $T(\boldsymbol{p})$ outputs larger values for the joint distribution of multiple independent random variables than for the marginal distributions of the individual random variables. However, while the entropy of the joint probability distribution is the sum of the entropies of the marginal probability distributions of the independent random variables [6], this property does not hold for $T(\boldsymbol{p})$ for finite n. This is the main difference between the entropy function and the polynomial approximation function $T(\boldsymbol{p})$, and justifies the use of the entropy function as a measure of uncertainty about a random variable for applications where the property of additivity is important.

IV. An unbiased estimator of the function $T(\boldsymbol{p})$

The problem is formulated as finding an unbiased and asymptotically convergent estimator of $T(\boldsymbol{p})$ with controlled variance. We will show that as far as estimation of $T(\boldsymbol{p})$ is concerned, the 'bias problem' can be solved completely, and does not require the specification of a prior distribution on the entropy.

We first consider the problem of deriving an unbiased estimate of the sum $S(\boldsymbol{p}, k) = \sum_{m=1}^{M} p_m^k$ for all $k \in \{1, \dots, n\}$. Let

$$\hat{S}(\boldsymbol{n}, k) = \sum_{m=1}^{M} c(n_m, n, k), \tag{6}$$

where

$$n_m \geq k \quad \Rightarrow \quad c(n_m, n, k) = \frac{n_m!(n-k)!}{n!(n_m-k)!},$$
$$n_m < k \quad \Rightarrow \quad c(n_m, n, k) = 0.$$

Since $\binom{n}{n_m}\binom{n_m}{k} = \binom{n}{k}\binom{n-k}{n_m-k}$ for all $n_k \geq k$,

$$\begin{aligned}
\mathrm{E}\{\hat{S}\} &= \sum_{m=1}^{M} \binom{n}{k}^{-1} \sum_{n_m=k}^{n} \Pr\{N_m = n_m\} \binom{n_m}{k} \\
&= \sum_{m=1}^{M} \binom{n}{k}^{-1} \sum_{n_m=k}^{n} \binom{n}{n_m} p_m^{n_m} (1-p_m)^{n-n_m} \binom{n_m}{k} \\
&= \sum_{m=1}^{M} p_m^k,
\end{aligned} \tag{7}$$

and the claim that the estimate is unbiased follows.

By the linearity of the expectation, it then follows that the unbiased estimator of $T(\boldsymbol{p})$ is given by

$$\hat{T}(\boldsymbol{n}) \equiv \sum_{k=1}^{n} a_k \hat{S}(\boldsymbol{n}, k)\,. \tag{8}$$

Using the Newton series for the digamma function $\psi(n_m) = -\gamma - \sum_{k=1}^{n_m} \frac{(-1)^k}{k} \binom{n_m}{k} - \frac{1}{n_m}$, the expression for the polynomial estimator then simplifies to

$$\hat{T}(\boldsymbol{n}) = \psi(n) - \frac{1}{n} \sum_{m=1}^{M} n_m \psi(n_m)\,. \tag{9}$$

It follows that $\hat{T}(\boldsymbol{n})$ is an asymptotically consistent estimator of $H(\boldsymbol{p})$, since $\psi(n_m) \to \ln(np_m)$ and $\psi(n) \to \ln(n)$ as $n \to \infty$. The asymptotic expansion series of $\psi(x)$ is given as $\psi(x) = \ln(x) - \frac{1}{2x} + \sum_{j=1}^{\infty} \frac{\zeta(1-2j)}{x^{2j}}$ with ζ the Riemann Zeta function. It follows that our estimator is closely related to the classic Miller-Madow estimator [7], which is known to have a faster convergence rate than the plugin estimator, and is defined as

$$\hat{M}(\boldsymbol{n}) \equiv \ln(n) - \frac{1}{2n} - \frac{1}{n} \sum_{m=1}^{M} n_m \left(\ln(n_m) - \frac{1}{2n_m} \right) \tag{10}$$

where we assume $\frac{0}{0} \equiv 0$. Our estimator is less biased than the Miller-Madow and plugin estimator. Since the inequality

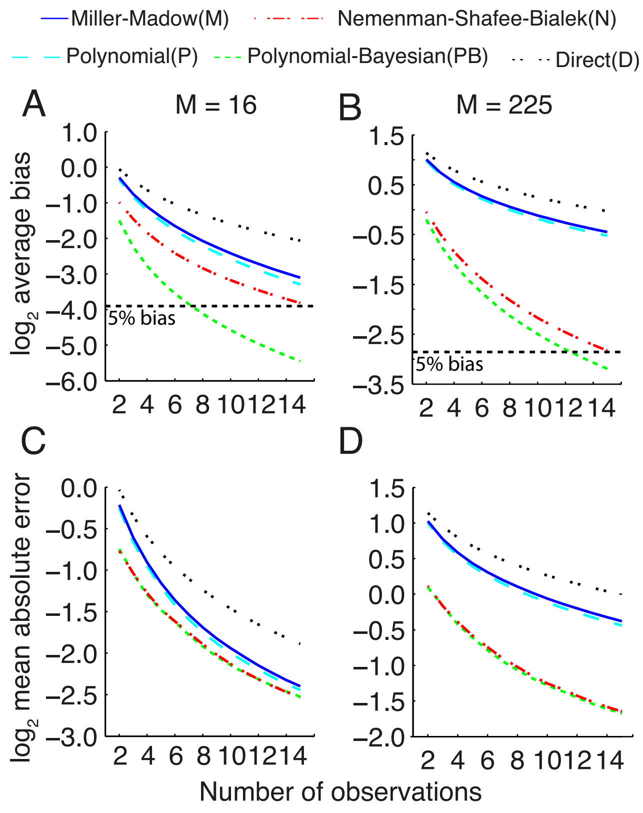

Fig. 1. Bias and average mean absolute error of various estimators. (A) Average $\log_2$ transformed bias (in nats) across the $[0, \log(M)]$ interval (*y-axis*), as a function of the number of observations n (*x-axis*), with $M = 16$. Dashed green: polynomial-Bayesian $\Upsilon_{\overline{\beta}}(\boldsymbol{p})$ entropy estimator. Dashed cyan: polynomial estimator $\hat{T}(\boldsymbol{n})$. Dashed-dotted red: NSB estimator. Blue: Miller-Madow estimator. Dashed-dotted black: plugin estimator. Dashed grey, horizontal: 5% bias of the average entropy across the $[0, \ln(M)]$ interval. (B) As in (A), but now for $M = 225$. (C) Average absolute error of the entropy estimates, averaged across the $[0, \log(M)]$ interval, with $M = 16$. Legends are similar to (A-B). (D): Similar to (C), but now for $M = 225$.

$\ln(n) - \frac{1}{2n} - \psi(n) \leq \ln(n_m) - \frac{1}{2n_m} - \psi(n_m)$ holds, it follows that $\hat{T}(\boldsymbol{n}) \geq \hat{M}(\boldsymbol{n}) \geq \hat{H}_{\text{plugin}}(\boldsymbol{n})$. As $\mathrm{E}\{\hat{T}(\boldsymbol{n})\} \leq H(\boldsymbol{p})$, the claim follows.

V. Estimation of the remainder function $R(\boldsymbol{p})$

Let us address the problem of constructing a Bayesian estimator of the remainder function $R(\boldsymbol{p})$. We first design such an estimator using a symmetric Dirichlet prior over the vectors $\boldsymbol{p}$, which has a disadvantage that it imposes an informative prior on $H(\boldsymbol{p})$ [3] (see Section VI). The symmetric Dirichlet distribution has a probability density function defined as $D_\beta(\boldsymbol{p}) = \Gamma(\beta)^M / \Gamma(M\beta) \prod_{m=1}^{M} p_m^{\beta-1}$. If $\beta \to 0$, then it is concentrated on probability vectors that specify small values of the entropy. If $\beta = 1$, then all $\boldsymbol{p} \in \Delta^M \equiv \{(p_1, \ldots, p_M) \in \mathcal{R}^M \mid (\sum_{m=1}^{M} p_m) = 1, p_m \geq 0 \,\text{for all}\, m\}$ have a probability density of 1, meaning that the prior on $\boldsymbol{p}$ is uniform.

We define the estimator of the entropy with Bayesian estimation of the remainder according to a Dirichlet prior as

$$\Upsilon_\beta(\boldsymbol{n}) \equiv \hat{T}(\boldsymbol{n}) + \hat{R}_\beta(\boldsymbol{n}) \,, \tag{11}$$

where $\hat{T}(\boldsymbol{n})$ is defined in Eq. 8 and $\hat{R}_\beta$ is the Bayesian estimator of the remainder based on a symmetric Dirichlet prior on $\boldsymbol{p}$ with concentration parameter β, defined as

$$\hat{R}_\beta(\boldsymbol{n}) \equiv \hat{H}_\beta(\boldsymbol{n}) - \hat{T}_\beta(\boldsymbol{n}) \,. \tag{12}$$

Here, $\hat{T}_\beta(\boldsymbol{n})$ is the Bayesian estimate of $T(\boldsymbol{p})$,

$$\hat{T}_\beta(\boldsymbol{n}) \equiv \sum_{m=1}^{M} \sum_{k=1}^{n} a_k \frac{\Gamma(\beta M + n)\Gamma(\beta + k + n_m)}{\Gamma(\beta M + n + k)\Gamma(\beta + n_m)} \tag{13}$$

and $\hat{H}_\beta(\boldsymbol{n})$ is the Bayesian estimate of the entropy,

$$\hat{H}_\beta(\boldsymbol{n}) \equiv \sum_{m=1}^{M} \frac{(n_m + \beta)}{(n + \beta M)} \cdot \\ (\psi(n + 1 + \beta M) - \psi(n_m + \beta + 1)) \,. \tag{14}$$

These expressions are obtained using the following considerations. By Bayes' theorem, the posterior probability of the vector $\boldsymbol{p}$ given observation of $\boldsymbol{n}$ and prior $P(\boldsymbol{p})$ is given as $P(\boldsymbol{p}|\boldsymbol{n}) = P(\boldsymbol{n}|\boldsymbol{p})P(\boldsymbol{p})/P(\boldsymbol{n})$ where P is a probability density function on the probability distribution of the source. The Bayesian estimator of the function of the probability distribution of the source, $Q(\boldsymbol{p})$, is then defined as $\hat{Q}(\boldsymbol{n}) \equiv \int Q(\boldsymbol{p})P(\boldsymbol{p}|\boldsymbol{n})d\boldsymbol{p}$. Wolpert and Wolf [4] have shown that the Bayesian estimator $\hat{Q}(\boldsymbol{n})$ with $P(\boldsymbol{p}) = D_\beta(\boldsymbol{p})$ is found by computing the ratio of integrals [4]

$$\hat{Q}(\boldsymbol{n}) = I[Q(\boldsymbol{p}), \boldsymbol{n}] / I[1, \boldsymbol{n}] \,, \tag{15}$$

where

$$I[Q(\boldsymbol{p}), \boldsymbol{n}] = \int Q(\boldsymbol{p}) \prod_{m=1}^{M} p_m^{n_m} \prod_{m=1}^{M} p_m^{\beta-1} d\boldsymbol{p} \,. \tag{16}$$

Note that $\prod_{m=1}^{M} p_m^{\beta-1} \propto D_\beta(\boldsymbol{p})$ where the beta normalization has been omitted because it is eliminated by division in eq. 15, and $\prod_{m=1}^{M} p_m^{n_m}$ is the multinomial probability of observing $\boldsymbol{n}$, where the multinomial normalization coefficient has been omitted again. Wolpert and Wolf [4] have shown that

$$I[1, \boldsymbol{n}] = \frac{\prod_{m=1}^{M} \Gamma(\beta + n_m)}{\Gamma\left(\sum_{m=1}^{M} (\beta + n_m)\right)} \,. \tag{17}$$

Using the extensive derivations from Section IV in [4] based on the Laplace transformation of $p_m^{n_m}$ (which hold similarly for $p_m^{n_m + \beta + 1}$), the expression for the Bayesian estimator of the entropy $\hat{H}_\beta(\boldsymbol{n})$ (eq. 14) given a symmetric Dirichlet prior is obtained.

Using the linearity property of integration, eq. 17, and the equality $\Gamma(x + y) = \Gamma(x)(x + y - 1)!/(x - 1)!$ for any (x, y), the integral $I[T(\boldsymbol{p}), \boldsymbol{n}]$ evaluates to

$$
\begin{aligned}
I[T(\boldsymbol{p}), \boldsymbol{n}] &= \sum_{k=1}^{n} a_k \int \prod_{m=1}^{M} p_{n_m}^{n_m+\beta-1} \sum_{m=1}^{M} p_m^k \, d\boldsymbol{p} \\
&= \sum_{k=0}^{n} a_k \sum_{l=1}^{M} \frac{\prod_{m=1}^{M} \Gamma\left(n_m + \beta + k\delta_{l,m}\right)}{\Gamma(n + \beta M + k)} \qquad (18) \\
&= \sum_{k=1}^{n} a_k \sum_{m=1}^{M} \frac{(n_m + \beta + k - 1)!}{(n_m + \beta - 1)!} \frac{\prod_{m=1}^{M} \Gamma(\beta + n_m)}{\Gamma(\beta M + n + k)} ,
\end{aligned}
$$

where $\delta_{l.m}$ is the Kronecker delta, and the $n_m + \beta$ terms in eq. 17 were replaced by $n_m + \beta + k\delta_{l,m}$. The expression for the Bayesian estimator $\hat{T}_\beta(\boldsymbol{n})$ in eq. 13 then follows.

VI. Bayesian Estimation of the Remainder Function Based on a Nearly Flat Prior on the Entropy

As discussed in the introduction, Nemenman et al. developed a nearly flat prior, as a mixture of Dirichlet distributions, on the entropy and a Bayesian (the NSB) entropy estimate based on that [3]. Here, we will use the same prior as developed by Nemenman et al. [3] to provide an estimator of the remainder $R(\boldsymbol{p})$. The uniform prior on the entropy has the following justification: When estimating the entropy without any *a priori* knowledge about the probability distribution of $\boldsymbol{p}$ or $H(\boldsymbol{p})$, we can either choose to use a uniform prior on $H(\boldsymbol{p})$ [3], or on $\boldsymbol{p}$ [4]. A uniform prior on $\boldsymbol{p}$ imposes an informative prior on $H(\boldsymbol{p})$, and vice versa. While the uniform prior on $\boldsymbol{p}$ is a sensible choice for computing the posterior estimate $P(\boldsymbol{p} \mid \boldsymbol{n})$, it dominates the estimation of $H(\boldsymbol{p})$, such that relatively small errors in estimating $\boldsymbol{p}$ are traded against relatively large errors in estimating $H(\boldsymbol{p})$.

Our polynomial-Bayesian estimator of the entropy $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ then becomes

$$\Upsilon_{\bar{\beta}}(\boldsymbol{n}) = \hat{T}(\boldsymbol{n}) + \hat{R}_{\bar{\beta}}(\boldsymbol{n}) \,. \qquad (19)$$

The Bayesian estimator of the remainder is defined as

$$\hat{R}_{\bar{\beta}}(\boldsymbol{n}) \equiv \frac{\int_0^\infty p(\beta, \boldsymbol{n}) \hat{R}_\beta(\boldsymbol{n}) \frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}{\int_0^\infty p(\beta, \boldsymbol{n}) \frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta} \,, \qquad (20)$$

where $\hat{R}_\beta(\boldsymbol{n})$ is defined in eq 12, and where the posterior probability density of β given observation of $\boldsymbol{n}$ is proportional to

$$p(\beta, \boldsymbol{n}) = \frac{\Gamma(M\beta)}{n + M\beta} \prod_{m=1}^{M} \frac{\Gamma(n_m + \beta)}{\Gamma(\beta)} \,, \qquad (21)$$

where $\mathrm{E}\{H(\boldsymbol{p});\beta\}$ is the expected value of the entropy for a Dirichlet distribution,

$$
\begin{aligned}
\mathrm{E}\{H(\boldsymbol{p});\beta\} &\equiv \int P_\beta(\boldsymbol{p}) H(\boldsymbol{p}) d\boldsymbol{p} \\
&= \psi_0(M\beta + 1) - \psi_0(\beta + 1) \,, \qquad (22)
\end{aligned}
$$

with $\psi_0(x)$ the polygamma function of order zero.

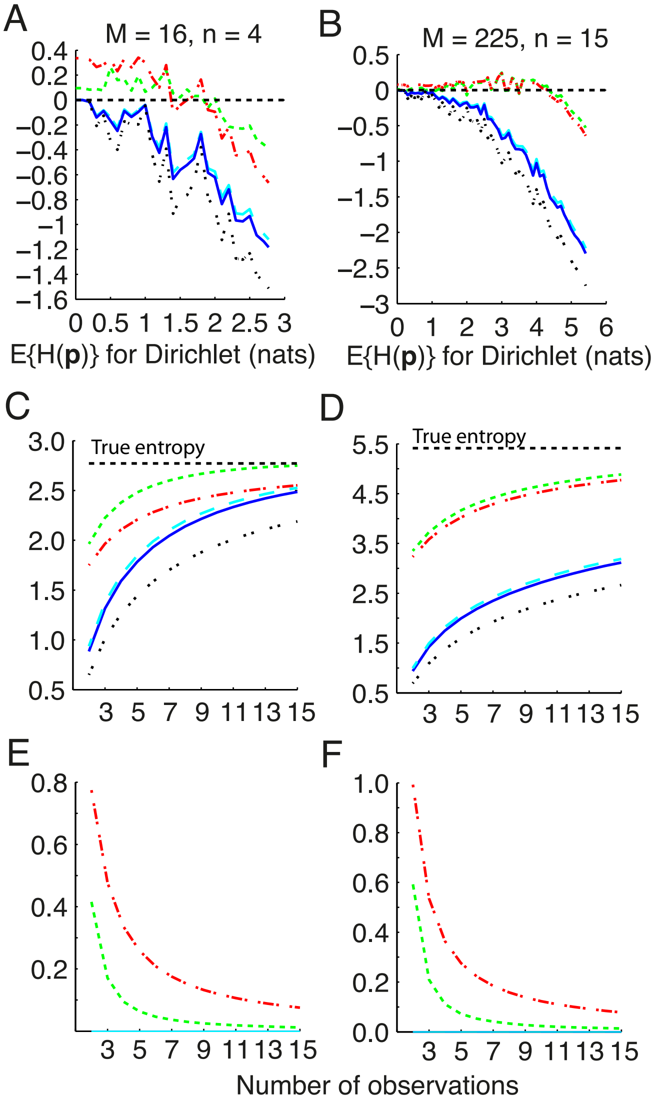

Fig. 2. (A-D) Expected entropy values for varying Dirichlet probability distributions. The expected entropy was parametrized by varying the concentration parameter β of a Dirichlet distribution and drawing a vector of probabilities $\boldsymbol{p}$ from this Dirichlet distribution. The x-axis corresponds to the expected entropy for a given Dirichlet distribution. The y-axis corresponds to the bias in nats. (A-B) correspond to $M = 16, 225$ and $n = 4, 15$ respectively. Color/line coding as in Figure 1. (C-D) Expected entropy of various estimators (*y-axis*) as a function of the number of observations with the expected entropy of the Dirichlet equal to $\ln(M) - 1/1000$. (C-D) correspond to $M = 16, 225$, respectively. (E-F) Similar to (C-D), but now for a Dirichlet distribution with the expected entropy equal to 1/10000.

The Bayesian estimator $\hat{R}_{\bar{\beta}}$ is based on a nearly flat prior distribution (the 'NSB' distribution) on the entropy that is obtained as mixture of Dirichlet distributions $P_\beta(\boldsymbol{p})$ [3] as

$$\bar{D}(\boldsymbol{p}) = \int_0^\infty P_\beta(\boldsymbol{p}) \frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta \,. \qquad (23)$$

The rationale of the Dirichlet mixture is that the distribution of the entropy $H(\boldsymbol{p})$ is very peaked around $\mathrm{E}\{H(\boldsymbol{p});\beta\}$ if $\boldsymbol{p}$ is Dirichlet-distributed with concentration parameter β [3]. Since the integral in eq. 23 runs effectively over $\frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta =$

$\mathrm{E}\{H(\boldsymbol{p});\beta\}$, it follows that the prior distribution $\bar{D}(\boldsymbol{p})$ of $\boldsymbol{p}$ translates into a nearly flat prior distribution of the entropy $H(\boldsymbol{p})$ for the interval $[0,\ln(M)]$ [3]. However, since there is some spread of $H(\boldsymbol{p})$ around $\mathrm{E}\{H(\boldsymbol{p});\beta\}$, the $\bar{D}(\boldsymbol{p})$ prior does not translate into a completely flat prior distribution of the entropy.

Nemenman et al. [3] showed that the expression for the Bayesian estimator of the entropy given this prior reduces to the one-dimensional integral

$$\begin{aligned}\hat{H}_{\bar{\beta}}(\boldsymbol{n}) &= \int H(\boldsymbol{p})P(\boldsymbol{p} \mid \boldsymbol{n})d\boldsymbol{p} \\ &= \frac{\int_0^\infty p(\beta,\boldsymbol{n})\hat{H}_\beta(\hat{\boldsymbol{p}})\frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta}d\beta}{\int_0^\infty p(\beta,\boldsymbol{n})\frac{d\,\mathrm{E}\{H(\boldsymbol{p});\beta\}}{d\beta}d\beta}. \qquad (24)\end{aligned}$$

Using the linearity property of integration, we substitute $\hat{H}_\beta(\boldsymbol{n})$ for $\hat{R}_\beta(\boldsymbol{n})$ and the definition eq. 20 is obtained.

VII. Comparison of bias and error of entropy estimators

We performed an exact (by computing probability of all $\boldsymbol{n}$) evaluation of the bias and mean absolute error of: (i) The plugin estimator $\hat{H}_{\text{plugin}}(\boldsymbol{n})$. (ii) The classic Miller-Madow estimator [7] (eq. 10) (iii) The Bayesian $\hat{H}_{\bar{\beta}}(\boldsymbol{n})$ estimator (eq. 24) [3]. (iv) Our newly developed polynomial $\hat{T}(\boldsymbol{n})$ (eq. 8) estimator and our polynomial-Bayesian estimator $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ (eq. 19). For a given number of symbols $M \in \{16, 225\}$, we varied the concentration parameter β of a symmetric Dirichlet distribution such that we covered the expected entropies in the interval $[1/10000, \ln(M) - 1/1000]$ with a step-size of 1/10. Note that the number of trials in neuroscience experiments is typically small and often does not exceed tens of trials per stimulus. For a particular symmetric Dirichlet distribution with concentration parameter β and expected entropy $\mathrm{E}\{H(\boldsymbol{p});\beta\}$ (eq. 22), we then drew a vector of probabilities $\boldsymbol{p}$ from that Dirichlet distribution. The entropy of the drawn vector of probabilities $\boldsymbol{p}$, $H(\boldsymbol{p})$, does not exactly coincide with $\mathrm{E}\{H(\boldsymbol{p});\beta\}$, but lies relatively close to it, since the distribution of $H(\boldsymbol{p})$ is highly peaked if $\boldsymbol{p}$ is Dirichlet-distributed [3].

Our polynomial-Bayesian $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ estimator, exhibited a significant reduction in bias in comparison to the other entropy estimators, having the lowest overall bias of all estimators, for all number of symbols M tested (Figure 1) ($M = 5$ and 81 gave similar results, data not presented). Furthermore, it had the smallest bias across the interval $[0, \ln(M)]$ for all estimators tested (Figure 2). The average reduction in bias in comparison to the NSB estimator was particularly strong when M was relatively small. This is a particularly useful property for applications where many symbols have $p_m = 0$, and M is over-estimated. Similarly, the reduction in bias of the $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ estimator relative to the NSB estimator was particularly pronounced when the entropy was close to 0, but was also present when the entropy was close to $\ln(M)$. This shows that, for this particular setting, the $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ estimator is, in terms of bias, more robust than the NSB estimator, since it exhibited the smallest maximum bias across probability vectors of $\boldsymbol{p}$ (minimax principle). However, the reduction in bias was accompanied by a balanced increase in variance such that the mean absolute errors of the NSB and $\Upsilon_{\bar{\beta}}(\boldsymbol{n})$ estimator were close to each other.

VIII. Conclusion

We have proposed a new algorithm for estimating the entropy, which is based on the representation of the entropy function as the sum of two polynomial terms, called the polynomial approximation function and the remainder. We have shown that an accurate and unbiased estimate of the value of the polynomial approximation function exists that does not depend on the choice of any particular prior on the probability distribution of the source. In addition, we have used the known Bayesian approach with a nearly flat prior on the entropy [3] to estimate the remainder. Our simulations show that the combined estimator essentially reduces the bias of the constructed estimate as compared to the known estimators. An advantage of our estimator relative to the NSB estimator is that part of the entropy function is estimated without the dependence on any particular prior, since the nearly uniform prior developed by [3] as a mixture of Dirichlet distributions is not the only possible uniform prior on the entropy, and is uniform only by approximation.

References

[1] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen, "Correcting for the sampling bias problem in spike train information measures," *J Neurophysiol*, vol. 98, pp. 1064–72, 2007.

[2] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[3] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Physical Review E*, vol. 69, p. 056111, 2004.

[4] D. H. Wolpert and D. R. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Phys Rev E*, vol. 52, pp. 6841–6854, 1995.

[5] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," *Arxiv preprint physics/0108025*, 2001.

[6] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[7] G. Miller, "Note on the bias on information estimates Information Theory in Psychology," in *Information Theory in Psychology; Problems and Methods II-B*, H. Quastler, Ed. Free Press, Glencoe, IL, 1955, pp. 95–100.