Estimation of the entropy based on its polynomial representation

Martin Vinck,[1] Francesco P. Battaglia,[1] Vladimir. B. Balakirsky,[2] A. J. Han Vinck,[2] and Cyriel M.A. Pennartz[1]

[1]*Cognitive and Systems Neuroscience Group, Center for Neuroscience, University of Amsterdam, the Netherlands*
[2]*Institüt für Experimentele Mathematik, Essen, Germany*

(Dated: April 20, 2012)

Estimating entropy from empirical samples of finite size is of central importance for information theory as well as the analysis of complex statistical systems. Yet, this delicate task is marred by intrinsic statistical bias. Here, we decompose the entropy function into a polynomial approximation function and a remainder function. The approximation function is based on a Taylor-expansion of the logarithm. Given n observations, we give an unbiased, linear estimate of the first n power series terms based on counting sets of k coincidences. For the remainder function, we use non-linear Bayesian estimation with a nearly flat prior distribution on the entropy that was developed by Nemenman, Shafee and Bialek. Our simulations show that the combined entropy estimator has reduced bias in comparison to other available estimators.

INTRODUCTION

Entropy and mutual information are of central importance for disciplines such as statistical mechanics and communication theory. 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 deal of attention in neurophysiology where the performance of networks of neurons is evaluated by information theoretical quantities [1–7]. In many of these applications, the probability distributions are not *a priori* known and have to be estimated from empirical data under limited sampling. For a physical system, this corresponds to knowledge of the phase space trajectory for a limited time interval [8]. Approximating the underlying (phase space) distribution with empirical frequency data 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 (e.g., see [9]). This is a serious issue for fields such as neuroscience, where obtaining very 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 approximate physiological state of learning, motivation, attention, etc. only for short periods.

To solve this problem, studies have attempted to evaluate the bias of the direct estimate of entropy and mutual information analytically [10–13], or they attempted at alleviating bias by reducing the dimensionality of the phase space [14–17]. 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 [18–20]. A promising framework to estimate entropy is the Bayesian one [21–23]. Wolpert and Wolf [23] developed a Bayesian estimator based on a uniform prior distribution of probabilities. A critical insight from Nemenman et al. [21, 22, 24] is that this uniform prior distribution of probabilities corresponds to a very peaked and dominating prior on the entropy with an expected value that lies relatively close to maximum entropy, causing large errors when the entropy is in fact small (see Fig 1 & 2). To solve this problem, Nemenman et al. [21, 22] developed a nearly flat prior distribution of the entropy and a Bayesian estimator, called the Nemenman-Shafee-Bialek (NSB) estimator, based on that. This nearly flat prior distribution is obtained as a mixture of symmetric Dirichlet distributions [21, 22]. The NSB estimator represents a major advance in entropy estimation, and has good performance in terms of bias and robustness in comparison to other available estimators [9, 21, 22, 25], as will be confirmed by our systematic comparisons between entropy estimators.

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 first obtain a polynomial representation of the entropy function, containing a polynomial approximation term and a remainder term. We show that an unbiased estimator exists for the polynomial approximation term. We then proceed by providing a Bayesian estimator of the remainder term, using eiter a flat prior on probabilities or a (near) flat prior on the entropy. We finish with a systematic and exact comparison between various available entropy estimators.

PROBLEM OF ENTROPY ESTIMATION

Suppose that a discrete memoryless source generates one of M values (states or symbols) from the set $\mathcal{R} = \{r_1, \ldots, r_M\}$ with probabilities $\boldsymbol{p} \equiv (p_1, \ldots, p_M)$. The (Shannon) entropy function [26] is defined as $H(\boldsymbol{p}) \equiv -\sum_{m=1}^{M} p_m \ln(p_m)$ (hereafter, we assume $0 \ln 0 = 0$). The vector of outcomes from 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 states r_m in the vector $\boldsymbol{x}$. The 'plugin' entropy estimator is defined as

$$\hat{H}_{\text{plugin}}(\boldsymbol{n}) = -\sum_{m=1}^{M} \frac{n_m}{n} \ln\left(\frac{n_m}{n}\right), \tag{1}$$

and it is well-known that it has a strict non–positive bias (see also Figure 1 and 2): As $-p_m \ln(p_m)$ is a concave function, 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 (e.g., see [18]).

The problem under consideration is to provide an estimate of $H(\boldsymbol{p})$ based on the observation $\boldsymbol{n}$, denoted $\hat{H}(\boldsymbol{n})$, 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, on average, small for all vectors of probabilities $\boldsymbol{p}$ (equivalently, for all entropies $H(\boldsymbol{p}) \in [0, \log(M)]$).

POLYNOMIAL REPRESENTATION OF THE ENTROPY FUNCTION.

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

$$H(\boldsymbol{p}) = T(\boldsymbol{p}) + R(\boldsymbol{p})\,. \tag{2}$$

We define the polynomial approximation $T(\boldsymbol{p})$ as

$$T(\boldsymbol{p}) \equiv a_1 + \sum_{m=1}^{M}\sum_{k=2}^{n} a_k p_m^k\,, \tag{3}$$

where the coefficients a_k are defined as

$$k = 1 \;\Rightarrow\; a_k = \sum_{j=1}^{n-1} 1/j\,, \tag{4}$$

$$k \geq 2 \;\Rightarrow\; a_k = \frac{(-1)^k(k-n-1)\binom{n}{n-k+1}}{(k-1)n}\,.$$

The expression for the remainder function $R(\boldsymbol{p})$ reduces to

$$R(\boldsymbol{p}) = \sum_{m=1}^{M} p_m(1-p_m)^n \sum_{k=0}^{\infty}\frac{(1-p_m)^k}{k+n}\,. \tag{5}$$

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 $-p_m\ln(p_m)$ 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

$$g(p_m) = \sum_{j=1}^{n-1}\frac{p_m}{j} + \sum_{k=1}^{n-1}\sum_{j=k}^{n-1}(-1)^k\,\frac{1}{j}\binom{j}{j-k}p_m^{k+1}\,. \tag{6}$$

The expressions for the coefficients and $T(\boldsymbol{p})$ then reduce to those in eqs. 3 and 4.

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 [26], this property does not hold for $T(\boldsymbol{p})$ for finite n (as is well-known, additivity is an important property of the entropy function as a measure of uncertainty [26]).

AN UNBIASED ESTIMATOR OF THE POLYNOMIAL APPROXIMATION FUNCTION

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 with a linear estimator, 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 $k \in \{2,\ldots,n\}$. Let

$$\hat{S}(\boldsymbol{n},k) = \sum_{m=1}^{M} c(n_m,n,k), \tag{7}$$

where

$$n_m \geq k \;\Rightarrow\; c(n_m,n,k) = \frac{n_m!(n-k)!}{n!(n_m-k)!}, \tag{8}$$

$$n_m < k \;\Rightarrow\; c(n_m,n,k) = 0.$$

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

$$\begin{aligned}\mathrm{E}\{\hat{S}\} &= \sum_{m=1}^{M}\binom{n}{k}^{-1}\sum_{n_m=k}^{n}\Pr\{N_m = n_m\}\binom{n_m}{k}\\ &= \sum_{m=1}^{M}\binom{n}{k}^{-1}\sum_{n_m=k}^{n}\binom{n}{n_m}p_m^{n_m}(1-p_m)^{n-n_m}\binom{n_m}{k}\\ &= \sum_{m=1}^{M}p_m^k\,,\end{aligned} \tag{9}$$

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

$$\hat{T}(\boldsymbol{n}) \equiv a_1 + \sum_{k=2}^{n} a_k\hat{S}(\boldsymbol{n},k)\,. \tag{10}$$

The structure of the presented estimator can be understood from the notion of counting coincidences: Note that $\binom{n_m}{k}$ corresponds to the number of k-tuples of observations that have the same outcome (i.e., coincidence), and that $\binom{n}{k}$ corresponds to the total number of k-tuples. Further note that the probability of observing the m-th state for k independent observations equals p_m^k, and that the probability of observing the same state for k independent observations equals $\sum_{m=1}^{M} p_m^k$.

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

$$\hat{T}(\boldsymbol{n}) = \psi(n) - \frac{1}{n}\sum_{m=1}^{M} n_m \psi(n_m)\,, \tag{11}$$

where $\psi(x)$ is the digamma function. Based on this expression, it can be seen that $\hat{T}(\boldsymbol{n})$ is an asymptotically consistent estimator of $H(\boldsymbol{p})$. As $n \rightarrow \infty$, $n_m \rightarrow np_m$, $\psi(n_m) \rightarrow \ln(np_m)$ and $\psi(n) \rightarrow \ln(n)$. Hence, the claim follows.

The expression has several relationships to other estimators. Firstly, it can be seen that it replaces the expression for the logarithm in the plugin-estimator by the digamma function. Secondly, estimator is related to the Grassberger estimator (eq. 23 in [13]), and differs only by a factor $\frac{1}{2n}$ as $\psi(x) \approx \ln(x) - \frac{1}{2x}$. Note that the Grassberger estimator contains additional correction factors that are based on the assumption that the symbol counts are Poisson distributed; this causes it to be improved for smaller p_m (see Figure 1 and 2). Thirdly, based on this approximation, it is seen directly that the estimator is related to the Miller-Madow correction, $\hat{H}(\boldsymbol{n}) + (L-1)/2n$, where L is defined as the estimated number of elements of the probability $\boldsymbol{p}$ that exceed zero. However, the estimator is slightly sharper, as it includes correction factors of higher power orders of $\frac{1}{n}$ as well (see Figure 1 and 2).

A COMBINED ESTIMATOR WITH BAYESIAN ESTIMATION OF THE REMAINDER FUNCTION

Let us address the problem of constructing a Bayesian estimator of the remainder function $R(\boldsymbol{p})$. By Bayes' theorem, the posterior probability of the vector $\boldsymbol{p}$ given observation of $\boldsymbol{n}$ and prior $P(\boldsymbol{p})$ is given as

$$P(\boldsymbol{p}|\boldsymbol{n}) = \frac{P(\boldsymbol{n}|\boldsymbol{p})P(\boldsymbol{p})}{P(\boldsymbol{n})}\,, \tag{12}$$

where $P(\boldsymbol{n}) = \int P(\boldsymbol{n}|\boldsymbol{p})P(\boldsymbol{p})d\boldsymbol{p}$.

The Bayesian estimator of a function $Q(\boldsymbol{p})$ is then defined as [23]

$$\hat{Q}_{\mathrm{Bay}}(\boldsymbol{n}) \equiv \int Q(\boldsymbol{p})P(\boldsymbol{p}|\boldsymbol{n})d\boldsymbol{p}\,. \tag{13}$$

We define our 'combined' estimator of the entropy with Bayesian estimation of the remainder as

$$\hat{H}_{\mathrm{comb}}(\boldsymbol{n}) \equiv \hat{T}(\boldsymbol{n}) + \hat{R}_{\mathrm{Bay}}(\boldsymbol{n})\,, \tag{14}$$

where $\hat{R}_{\mathrm{Bay}}(\boldsymbol{n})$ is a Bayesian estimator of the remainder function $R(\boldsymbol{p})$, with

$$\begin{aligned}\hat{R}_{\mathrm{Bay}}(\boldsymbol{n}) &\equiv \int \left(H(\boldsymbol{p}) - T(\boldsymbol{p})\right) P(\boldsymbol{p}|\boldsymbol{n})d\boldsymbol{p} \\ &\equiv \hat{H}_{\mathrm{Bay}}(\boldsymbol{n}) - \hat{T}_{\mathrm{Bay}}(\boldsymbol{n})\,.\end{aligned} \tag{15}$$

Bayesian estimation with a symmetric Dirichlet prior

We first consider Bayesian estimation with a symmetric Dirichlet prior, the known conjugate prior of the multinomial distribution, that has a probability density function defined as

$$D_\beta(\boldsymbol{p}) = \frac{1}{B(\boldsymbol{\beta})}\prod_{m=1}^{M} p_m^{\beta-1}\,, \tag{16}$$

where

$$B(\boldsymbol{t}) \equiv \frac{\prod_{m=1}^{M}\Gamma(t_m)}{\Gamma\left(\sum_{m=1}^{M} t_m\right)} \tag{17}$$

is the multinomial beta function for any vector $\boldsymbol{t} \geq \boldsymbol{0}$ with number of elements M, such that $\int \prod_{m=1}^{M} p_m^{t_m} d\boldsymbol{p} = B(\boldsymbol{t})$. For $\beta \rightarrow 0$ and $\beta \rightarrow \infty$, the Dirichlet distribution is concentrated on probability vectors $\boldsymbol{p}$ that specify minimum and maximum values of the entropy, respectively. If $\beta = 1$, then the prior on $\boldsymbol{p}$ is uniform.

Let the prior $P(\boldsymbol{p}) = D_\beta(\boldsymbol{p})$ (the disadvantages of this prior will be discussed below). The maximum likelihood estimate of the entropy under this prior equals the plugin estimate $\hat{H}_{\mathrm{plugin}}$ (eq. 1), since $\frac{\boldsymbol{n}}{n} = \mathrm{argmax}_{\boldsymbol{p}}\, P(\boldsymbol{p}|\boldsymbol{n})$ and $\hat{H}_{\mathrm{plugin}} = H(\frac{\boldsymbol{n}}{n})$. The Bayesian estimate of the entropy under this prior has the known expression (see Section IV in [23])

$$\hat{H}^{\mathrm{Dir}}_{\mathrm{Bay}} = \psi_0(n+1+\beta M) - \sum_{m=1}^{M}\frac{n_m+\beta}{n+\beta M}\psi_0(n_m+\beta+1) \tag{18}$$

with $\psi_0(x)$ the digamma function. Note that the plugin entropy estimate can be written in similar functional form as $\ln(n) - \sum_{m=1}^{M}\frac{n_m}{n}\ln(n_m)$, and that $\lim_{x\rightarrow\infty}\psi_0(x) = \ln(x)$. Hence, the conclusion follows that as $n \rightarrow \infty$, $\hat{H}^{\mathrm{Dir}}_{\mathrm{Bay}}(\boldsymbol{n}) \rightarrow \hat{H}_{\mathrm{plugin}}(\boldsymbol{n}) \rightarrow H(\boldsymbol{p})$.

We derive the expression for the Bayesian estimate of $T(\boldsymbol{p})$ as

$$\hat{T}^{\mathrm{Dir}}_{\mathrm{Bay}}(\boldsymbol{n}) = \sum_{m=1}^{M}\sum_{k=1}^{n} a_k \frac{\Gamma(\beta M+n)\Gamma(\beta+k+n_m)}{\Gamma(\beta M+n+k)\Gamma(\beta+n_m)}\,, \tag{19}$$

where the coefficients a_k are defined in eq. 4. Since the multinomial likelihood equals $P(\boldsymbol{n}|\boldsymbol{p}) = \prod_{m=1}^{M} p_m^{n_m}/B(1+\boldsymbol{n})$, and $P(\boldsymbol{n}) = \frac{B(\boldsymbol{\beta}+\boldsymbol{n})}{B(\boldsymbol{\beta})B(1+\boldsymbol{n})}$, the expressions of eq. 18 and 19 are obtained by solving the integral

$$\begin{aligned}\hat{Q}_{\text{Bay}}(\boldsymbol{n}) &\equiv \int Q(\boldsymbol{p})P(\boldsymbol{p}|\boldsymbol{n})d\boldsymbol{p}\\ &= \frac{\int Q(\boldsymbol{p}) \prod_{m=1}^{M} p_m^{n_m} \prod_{m=1}^{M} p_m^{\beta-1} d\boldsymbol{p}}{\int \prod_{m=1}^{M} p_m^{n_m} \prod_{m=1}^{M} p_m^{\beta-1} d\boldsymbol{p}} \\ &\equiv \frac{I[Q(\boldsymbol{p}), \boldsymbol{n}]}{B(\beta + \boldsymbol{n})}\,.\end{aligned} \tag{20}$$

Using the linearity property of integration, the integral $I[T(\boldsymbol{p}), \boldsymbol{n}]$ evaluates to

$$\begin{aligned}I[T(\boldsymbol{p}), \boldsymbol{n}] &= \sum_{k=1}^{n} a_k \int \prod_{m=1}^{M} p_{n_m}^{n_m+\beta-1} \sum_{m=1}^{M} p_m^k \, d\boldsymbol{p}\\ &= \sum_{k=1}^{n} a_k \sum_{l=1}^{M} \frac{\Gamma(\beta + n_l + k) \prod_{m=1}^{M} \Gamma(\beta + n_m)}{\Gamma(\beta + n_l)\Gamma(M\beta + n + k)}\,.\end{aligned} \tag{21}$$

By combining eqs. 20 and 21, the expression for the estimator in eq. 19 follows. The combined estimator with Dirichlet prior $D_\beta(\boldsymbol{p})$ is then defined as

$$\hat{H}^{\text{Dir}}_{\text{comb}} = \hat{T}(\boldsymbol{n}) + \hat{H}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n}) - \hat{T}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n})\,. \tag{22}$$

Bayesian estimation of the remainder function based on the nearly flat NSB 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 (NSB) entropy estimate based on this prior [21]. Here, we will use the same prior as developed by Nemenman et al. [21] to provide an estimator of the remainder $R(\boldsymbol{p})$. The near uniform prior on the entropy attempts to maximize the prior uncertainty (entropy) about the uncertainty (entropy), and its use has the following rationale: When estimating the entropy without any *a priori* knowledge about the probability distribution of the probability vector $\boldsymbol{p}$ or $H(\boldsymbol{p})$, the two reasonable choices at hand would be to use a (near) uniform prior on $H(\boldsymbol{p})$ [21], or on $\boldsymbol{p}$ [23]. A uniform prior on $\boldsymbol{p}$ imposes a very informative prior on $H(\boldsymbol{p})$ [22], and vice versa. While the uniform prior on $\boldsymbol{p}$ is a sensible choice for computing the posterior estimate $P(\boldsymbol{p}|\boldsymbol{n})$, 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 $H(\boldsymbol{p})$ (see Figure 1 and 2).

The NSB distribution is defined as mixture of Dirichlet distributions [21],

$$D_{\text{NSB}}(\boldsymbol{p}) = \int_0^\infty \frac{1}{\ln(M)B(\boldsymbol{\beta})} \prod_{m=1}^{M} p_m^{\beta-1} \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta\,. \tag{23}$$

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

$$\begin{aligned}\text{E}\{H(\boldsymbol{p});\beta\} &\equiv \int D_\beta(\boldsymbol{p})H(\boldsymbol{p})d\boldsymbol{p}\\ &= \psi_0(M\beta + 1) - \psi_0(\beta + 1)\,,\end{aligned} \tag{24}$$

as can be seen by letting $\boldsymbol{n} = \boldsymbol{0}$ in eq. 18. The normalization factor $\frac{1}{\ln(M)}$ ensures that $\int D_{\text{NSB}}(\boldsymbol{p})d\boldsymbol{p} = 1$.

The rationale of the Dirichlet mixture is that the distribution of the entropy $H(\boldsymbol{p})$ is very peaked around $\text{E}\{H(\boldsymbol{p});\beta\}$ if the probability vector $\boldsymbol{p}$ is Dirichlet-distributed with concentration parameter β [21]. Since the integral in eq. 23 runs effectively over $\frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta}d\beta = \text{E}\{H(\boldsymbol{p});\beta\}$, it follows that the prior distribution $D_{\text{NSB}}(\boldsymbol{p})$ translates into a nearly flat prior distribution of the entropy $H(\boldsymbol{p})$ for the interval $[0, \ln(M)]$ nats [21]. However, since there is some spread in the distribution of $H(\boldsymbol{p})$ around $\text{E}\{H(\boldsymbol{p});\beta\}$, which becomes especially skewed around $\beta = 0$, the $D_{\text{NSB}}(\boldsymbol{p})$ prior does not translate into a completely flat prior distribution of the entropy.

Let $P(\boldsymbol{p}) = D_{\text{NSB}}(\boldsymbol{p})$. The known expression for the Bayesian estimator of the entropy under the NSB prior [21] then equals

$$\begin{aligned}\hat{H}^{\text{NSB}}_{\text{Bay}}(\boldsymbol{n}) &= \int H(\boldsymbol{p})P(\boldsymbol{p}|\boldsymbol{n})d\boldsymbol{p}\\ &= \frac{\int_0^\infty \left(\int H(\boldsymbol{p})P_\beta(\boldsymbol{p}) \prod_{m=1}^{M} p^{n_m} d\boldsymbol{p}\right) \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}{\int_0^\infty \left(\int P_\beta(\boldsymbol{p}) \prod_{m=1}^{M} p^{n_m} d\boldsymbol{p}\right) \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}\\ &= \frac{\int_0^\infty \frac{B(\beta+\boldsymbol{n})}{B(\boldsymbol{\beta})} \hat{H}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n}) \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}{\int_0^\infty \frac{B(\beta+\boldsymbol{n})}{B(\boldsymbol{\beta})} \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}\,.\end{aligned} \tag{25}$$

We define the Bayesian estimator of the polynomial approximation function as

$$\hat{T}^{\text{NSB}}_{\text{Bay}}(\boldsymbol{n}) = \frac{\int_0^\infty \frac{B(\beta+\boldsymbol{n})}{B(\boldsymbol{\beta})} \hat{T}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n}) \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}{\int_0^\infty \frac{B(\beta+\boldsymbol{n})}{B(\boldsymbol{\beta})} \frac{d\,\text{E}\{H(\boldsymbol{p});\beta\}}{d\beta} d\beta}\,. \tag{26}$$

The combined estimator with NSB prior is then defined as

$$\hat{H}_{\text{comb}}(\boldsymbol{n}) = \hat{T}(\boldsymbol{n}) + \hat{H}^{\text{NSB}}_{\text{Bay}}(\boldsymbol{n}) - \hat{T}^{\text{NSB}}_{\text{Bay}}(\boldsymbol{n})\,. \tag{27}$$

The integrals in eq. 25 and 27 have no known analytic solution and must be integrated numerically. To integrate across the complete interval $[0, \infty]$, we defined $\beta = \frac{1-t}{t}$, substituted $d\beta = \frac{dt}{t^2}$ and changed the integration limits to $t \in [0, 1]$.

It is important to note that the D_{NSB} prior may not be the ‘best’ flat prior on the entropy. The ‘best’ flat prior on the entropy should ideally assign equal prior probabilities to probability vectors $\boldsymbol{p}$ that have equal entropies, such that $P(\boldsymbol{p}|H(\boldsymbol{p})) = c$ where c is a constant, i.e. should not impose more prior information on $\boldsymbol{p}$ than is required to make the prior distribution of $H(\boldsymbol{p})$ flat. We give a counterexample to show that this constraint is not satisfied. Consider the vectors $\boldsymbol{p}_1 = (0.0013, 0.2310, 0.7677)$ and $\boldsymbol{p}_2 = (0.0892, 0.8396, 0.0712)$. We have $H(\boldsymbol{p}_1) = 0.550$ and $H(\boldsymbol{p}_2) = 0.550$. However, $D_{\text{NSB}}(\boldsymbol{p}_1) = 5.97$ and $D_{\text{NSB}}(\boldsymbol{p}_2) = 1.00$. Thus, for this particular example, the probabilities assigned to two vectors $\boldsymbol{p}$ that have equal entropies differ by a factor 6.

COMPARISON OF BIAS AND ERROR OF ENTROPY ESTIMATORS

In this section, we perform a systematic and exact evaluation of the bias, i.e., $|\,\mathrm{E}\{\hat{H}(\boldsymbol{n})\} - H(\boldsymbol{p})|$ where $\hat{H}$ is an estimator of the entropy H, and mean absolute error, i.e. $\mathrm{E}\{|\hat{H} - H|\}$. This is done for several estimators: The plugin estimator $\hat{H}_{\text{plugin}}(\boldsymbol{n})$ (eq. 1); the classical Miller-Madow estimator [12], which is often taken as a benchmark in entropy estimation, and the related Panzeri-Treves estimator [10, 27, 28]; the Grassberger estimator [13, 20, 29], a well-known estimator from the linear (see below) class of estimators that improves on the Miller-Madow estimator; the Bayesian NSB estimator $\hat{H}^{\text{NSB}}_{\text{Bay}}(\boldsymbol{n})$ (eq. 25) [21, 22]; the Bayesian estimator $\hat{H}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n})$ with a flat prior on the probability vector $\boldsymbol{p}$ (i.e., $\beta = 1$) [23]; and, finally, our newly developed polynomial $\hat{T}(\boldsymbol{n})$ (eq. 10) estimator and the combined estimator $\hat{H}_{\text{comb}}(\boldsymbol{n})$ that uses the NSB prior (eq. 27).

All simulations were performed using Mathematica. For the NSB and Panzeri-Treves estimators, results using Mathematica were compared with those obtained by using the MATLAB code developed by the authors of [21] (http://nsb-entropy.sourceforge.net/) and the information breakdown toolbox [28], respectively.

Overview of various estimators

The well-known Miller-Madow [12] and related Panzeri-Treves [27] estimator is based on a Taylor expansion of the bias of the plugin-estimator of the entropy, and is defined as $\hat{H}(\boldsymbol{n}) + (L-1)/2n$, where L is defined as the estimated number of elements of the probability $\boldsymbol{p}$ that exceed zero. The classic Miller-Madow estimator is defined by setting $L = |\{m : n_m > 0\}|$, the empirically observed number of non-zero elements in the vector of observations $\boldsymbol{n}$. In addition, we evaluated the Panzeri-Treves estimator [27], which is based on a Bayesian estimate of L, where we set the constraint $L \leq M$.

The Grassberger estimator ([13, 20, 29]) is defined as $\ln(n) - \frac{1}{n}\sum_{m=1}^{M} n_m G(n_m)$ (eq. 35 in [29]), where Grassberger defined the function $G(n_m) \equiv \psi_0(n_m) + (-1)^{n_m}\int_0^1 \frac{x^{n_m-1}}{x+1}dx$. This estimator is based on modeling the observed symbol counts $\boldsymbol{n}$ as outcomes of a Poisson process. Again, convergence is obtained by noting that $G(x) \to \ln(x)$ as $x \to \infty$.

The various entropy estimators can be divided into two classes according to their functional form, namely a linear class and non-linear class. The Grassberger [13], $\hat{H}_{\text{plugin}}(\boldsymbol{n})$, Miller-Madow with $L = |\{m : n_m > 0\}|$ [12], and $\hat{T}(\boldsymbol{n})$ (eq. 10) estimator are linear, since they can be written in the functional form $\sum_{m=1}^{M} f(n_m, n)$. Since $f(0, n) = 0$ for all of these linear estimators, they do not depend on knowledge of the number of states M. The $\hat{H}^{\text{Dir}}_{\text{Bay}}(\boldsymbol{n})$ (eq. 18) estimator is also linear, however, the equality $f(0, n) = 0$ typically does not hold for this estimator, implying that knowledge of the number of states M is required.

The Panzeri-Treves [27], NSB [21] and our combined $\hat{H}_{\text{comb}}(\boldsymbol{n})$ estimator (eq. 27) have a non-linear form, i.e., they cannot be decomposed into the linear form $\sum_{m=1}^{M} f(n_m, n)$. Nevertheless, the Panzeri-Treves and the $\hat{H}_{\text{comb}}(\boldsymbol{n})$ estimator can be decomposed into a linear term and a non-linear component, while the NSB estimator is strictly non-linear, due to the function $B(\beta + \boldsymbol{n})$ in eq. 25. All of these 'non-linear' estimators also depend on knowledge of the number of states M.

Methodology of entropy evaluations

We performed a systematic evaluation of the various estimators by examining their average performance in terms of bias and mean absolute error across the $[0, \ln(M)]$ nats interval (Figure 1 and 2). This was achieved by, for a given number of states $M \in \{16, 81, 225\}$, varying the concentration parameter β of a symmetric Dirichlet distribution such that we covered the expected entropies in the interval $[1/10000, \ln(M) - 1/1000]$ nats with a step-size of 1/10 nats. For a particular symmetric Dirichlet distribution with concentration parameter β and expected entropy $\mathrm{E}\{H(\boldsymbol{p}); \beta\}$ (eq. 24), we then drew a probability test vector $\boldsymbol{p}$ 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 [22].

Both the linear and non-linear estimators that we discussed are symmetric in the sense that the particular order of the symbol counts $\boldsymbol{n}$ is irrelevant. We can describe $\boldsymbol{n}$ as a partition of the number of observations n. Knowledge of the ordered partition of the integer n is sufficient to determine the estimator output (the ordered partitions of the integer 4 with $M = 5$ are $\{4, 0, 0, 0, 0\}, \{3, 1, 0, 0, 0\}, \{2, 2, 0, 0, 0\}, \{2, 1, 1, 0, 0\}$ and $\{1, 1, 1, 1, 0\}$ for example). To obtain the exact values for the bias and mean absolute error of the various estimators, we therefore computed the probability of all partitions of the integer n according to a multinomial probability distribution with vector of probabilities $\boldsymbol{p}$ (the details of the employed algorithm will be subject of a separate work).

Comparison between the linear estimators

We first compared the performance of our polynomial $\hat{T}(\boldsymbol{n})$ estimator with the other linear estimators. The average bias and mean absolute error of the $\hat{T}(\boldsymbol{n})$ estimator were smaller than for the classic Miller-Madow estimator (Figure 1 & 2). We found that $\hat{T}(\boldsymbol{n})$ had higher and lower bias than the Grassberger estimator for large and small entropies, respectively (Figure 2). However, the average bias of $\hat{T}(\boldsymbol{n})$ was higher than for the Grassberger estimator (Figure 1). In fact, of all the tested estimators that do not require knowledge about the number of states M, the Grassberger estimator showed the best overall performance in terms of bias and mean absolute error for larger M. The difference between the Grassberger estimator and the $\hat{T}(\boldsymbol{n})$ estimator is a consequence of the fact that the Grassberger estimator is based on modeling the symbol counts (the n_m's) according to a Poisson distribution, which is a poor approximation when one of the p_m's is large, while the polynomial $\hat{T}(\boldsymbol{n})$ estimator is based on a Taylor-expansion of the logarithm around probability $p_m = 1$, which is a poor approximation if the p_m's are, on average, small.

Performance gain of the Bayesian estimators

The Bayesian NSB estimator [21, 22], using a nearly flat prior distribution of the entropy, had strongly reduced mean absolute error and overall bias for larger symbol count M (see $M = 81$ and $M = 225$) (Figure 1) in comparison to all the other available estimators (except in comparison to our $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimator, see below). This performance gain was caused by a smaller bias for entropies close to $\ln(M)$ nats, which outweighed the relative increase in bias of the NSB estimator for entropies close to 0 nats. However, for a smaller number of states ($M = 16$), the NSB estimator had an increased bias in comparison the Grassberger and Panzeri-Treves estimator. The Panzeri-Treves estimator improved strongly on the related Miller-Madow estimator and its performance was comparable to the Grassberger estimator, but it exhibited higher bias and mean absolute error than the NSB estimator when M was larger. The Bayesian estimator with a uniform prior distribution on $\boldsymbol{p}$ [23] had a very high overall bias and exhibited very slow convergence (Figure 1) when drawing probability test vectors $\boldsymbol{p}$ from the NSB prior.

Comparison between the NSB estimator and the combined estimator

Our combined $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimator exhibited a reduction in bias in comparison to the NSB estimator and had the lowest overall bias of all tested estimators, for all number of states M tested (Figure 1). The estimator required less than $n \approx \sqrt{M}$ observations to obtain an average bias of 10% of the average entropy across the $[0, \ln(M)]$ nats interval. In addition, as can be seen from Figure 2, it required about $n = \sqrt{M}$ observations to obtain a maximum bias of 10% across the complete $[0, \ln(M)]$ nats interval. As pointed out by Nemenman et al. [21], entropy estimators should be able to perform well in the $n \approx \sqrt{M}$ regime. This claim is based on the known fact that the number of observations required such that there is a probability p that the same outcome is observed for two observations (i.e., a coincidence) scales with $\sqrt{M}$ if all the p_m's are equal (the so called birthday problem), and on the notion that the occurrence of coincidences is critical to estimate entropy [8, 30] (see also eq. 9).

The average reduction in bias in comparison to the NSB estimator was particularly strong when M or the entropy was relatively small, since the $T(\boldsymbol{p})$ function is a particularly good approximation of $H(\boldsymbol{p})$ when the p_m's are relatively large. In contrast to the NSB estimator, the $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimator also outperformed the Grassberger, $\hat{T}(\boldsymbol{n})$ and Panzeri-Treves estimator in terms of bias for smaller M. The reduction in bias of the $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimator relative to the NSB estimator was particularly pronounced when the entropy was close to 0 nats, but was also present when the entropy was close to $\ln(M)$ nats.

However, the reduction in estimator bias was accompanied by a balanced increase in variance such that the mean absolute errors of the NSB and $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimator were about equal (but smaller than for the other tested estimators). For neurophysiological applications however, bias may often be more important than variance, since averaging entropy and mutual information estimates across independent measurements (stimuli, neurons, sessions, animals) will strongly reduce variance but not necessarily bias. However, an additional advantage of the NSB and $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ estimators is that their average (signed) bias lies close to zero (Figure 2), such that averaging estimates of varying entropies across independent experiments may further reduce estimator bias.

Minimization of the maximum bias

An alternative criterion by which we can evaluate entropy estimators independent of any prior is the maximum (across probability vectors $\boldsymbol{p}$) estimator risk (i.e., estimator bias or mean absolute error), also commonly referred to as the minimax principle. Our simulations demonstrate that the maximum bias of the NSB and our combined estimator across drawn probability vectors was minimal in comparison to the other tested estimators (Figure 2). The Bayesian estimator with flat prior on $\boldsymbol{p}$ on the other hand had the highest maximum risk of all the tested estimators, caused by the dominant, informative prior on the entropy $H(\boldsymbol{p})$ (Figure 2).

DISCUSSION

Our paper makes two contributions to the problem of entropy estimation. Firstly, 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 polynomial approximation function exists that does not depend on the choice of any particular prior on the probability distribution of the source, based on counting sets of k coincidences. This estimation procedure extends the work from Ma [8], who defined an entropy estimator based on counting sets of two coincidences. In addition, we have used the known Bayesian approach [21, 23] with a nearly flat prior on the entropy [21] to estimate the remainder. Our estimation strategy can be readily extended to the estimation of any function of a vector of probabilities $\boldsymbol{p}$. Secondly, we have performed simulations in which the bias and mean absolute error of estimators were computed exactly, and for probability vectors $\boldsymbol{p}$ whose entropies covered the whole $[0, \ln(M)]$ nats interval. Our simulations show that the NSB estimator [21] represents an important advance in entropy estimation as it strongly reduces the estimator bias relative to the other available estimators when the number of states is relatively large, and that our combined estimator further reduces the bias of the constructed estimate as compared to the known estimators (Figure 1 and 2).

Two critical points can be raised with respect to these results. Firstly, the nearly uniform NSB prior [21] is not the only possible (near) flat prior on the entropy, and is flat only by approximation, although it is the least informative prior on the entropy from the known prior distributions. We pointed out that probability vectors with the same entropy can still be assigned different prior probabilities according to the prior NSB distribution. This suggests that there may exist a better flat prior on the entropy than the NSB prior. An advantage of our combined estimator relative to the NSB estimator is that it estimates part of the entropy function without the dependence on any particular prior. In fact, our polynomial estimator is a 'universally' good estimator of the polynomial approximation function when taking bias as the loss function, since it is unbiased for any given vector of probabilities, just like the vector of empirical frequencies $\boldsymbol{n}/n$ is an unbiased estimator of any vector of probabilities $\boldsymbol{p}$. The NSB prior is then used only insofar as estimation of the remainder, for which no unbiased estimator is available, is concerned.

Secondly, it should be emphasized that our evaluation of average performance was inherently circular, in the sense that the estimators were evaluated for probability vectors that were drawn from the exact same prior (NSB) distribution [21] as used by the NSB estimator and our combined estimator. The notion of an 'average' performance of an estimator presupposes the choice of a prior assigning particular weights to different probability vectors $\boldsymbol{p}$. Thus, whether an estimator is on average 'good' does not depend only on the average performance under a chosen prior, but also on the question whether the chosen prior distribution is 'good'. All the analyzed estimators are not 'universally good', since for each estimator there exist probability test vectors $\boldsymbol{p}$ such that the estimators have a relatively poor performance; none of the examined estimators exhibit minimum risk (in terms of bias and mean absolute error) for all possible vectors $\boldsymbol{p}$. For example, the Bayesian estimator that is based on a flat prior on $\boldsymbol{p}$ [23] exhibits, on average, very good performance when evaluating the estimator for probability vectors drawn from this flat prior distribution. However, it estimates low entropies with a very large error, and also converges slowly when the entropy is close to maximum (Figure 2). On the other hand, the Miller-Madow [12], Panzeri-Treves [27] and our polynomial $\hat{T}(\boldsymbol{n})$ estimator show relatively large and small errors in the high and low entropy regime, respectively (Figure 2).

Nevertheless, there are two prominent reasons why Bayesian estimation with the NSB prior is to be preferred above Bayesian estimation with a uniform prior on the source probabilities $\boldsymbol{p}$. Firstly, while for many levels of the entropy the convergence of the Bayesian estimator with flat prior on $\boldsymbol{p}$ is extremely sluggish due to the dominant prior (Figure 1 and 2), the NSB and combined estimator have rapid convergence, also for entropies close to the expected entropy under the uniform prior on $\boldsymbol{p}$. This difference in performance is caused by the trade-off in estimating $\boldsymbol{p}$ and its entropy $H(\boldsymbol{p})$: The NSB prior produces large errors when estimating $\boldsymbol{p}$, but small errors when estimating $H(\boldsymbol{p})$. Conversely, the flat prior on $\boldsymbol{p}$ leads to small errors when estimating $\boldsymbol{p}$, but to large errors when estimating $H(\boldsymbol{p})$, unless the entropy is very close to the expected entropy under a flat prior on $\boldsymbol{p}$. Similarly, the empirically observed frequencies $\boldsymbol{n}/n$ provide an unbiased maximum likelihood estimate of $\boldsymbol{p}$, however $H(\boldsymbol{n}/n)$ is a very poor estimator of $H(\boldsymbol{p})$. Thus, paradoxically one needs to accept larger errors in estimating $\boldsymbol{p}$ in order to estimate $H(\boldsymbol{p})$ with smaller error. Secondly, the maximum bias of the NSB and combined estimator across drawn probability vectors was minimal in comparison to the other available estimators (minimax criterion), demonstrating their robustness. Conversely, the Bayesian estimator with uniform prior on $\boldsymbol{p}$ had the highest risk of all the estimators (for $H(\boldsymbol{p}) = 0$). The fact that our simulations evaluated the estimator performance across the $[0, \ln(M)]$ nats interval ensured that 'difficult' probability test vectors with highly variable entropies were drawn, which would not have occurred had probability vectors been drawn from a uniform prior distribution on $\boldsymbol{p}$ (since in that case the probability of observing small entropies would have been very low [22]). Similarly, Bayesian estimation with the nearly flat NSB prior on the entropy ensures that estimator performance does not break down for 'difficult' probability vectors, and correspondingly the NSB and our combined estimator have the smallest maximum bias of the available estimators, despite the fact that the Bayesian estimators minimize the average risk for a given prior, not the maximum risk.

ACKNOWLEDGEMENTS

We thank Dr. P. Gori-Giorgi for helpful comments on the manuscript. This work was supported by VICI grant 918.46.609 and EU grants FP7-217148 and FP7-ICT-270108.

[1] R. Quiroga and S. Panzeri, Nature Reviews Neuroscience, **10**, 173 (2009).

[2] F. Rieke, D. Warland, and W. Bialek, *Spikes: exploring the neural code* (The MIT Press, 1999).

[3] P. Dayan, L. Abbott, and L. Abbott, *Theoretical neuroscience: Computational and mathematical modeling of neural systems* (MIT Press, 2001).

[4] D. MacKay and W. McCulloch, Bulletin of Mathematical Biology, **14**, 127 (1952).

[5] J. Hertz and S. Panzeri, in *The Handbook of Brain Theory and Neural Networks*, edited by M. B. Arbib (MIT Press, 2002) pp. 1023–1026.

[6] A. Borst and F. Theunissen, Nature Neuroscience, **2**, 947 (1999).

[7] W. Bialek and F. Rieke, Trends in Neuroscience, **15**, 428 (1992).

[8] S. Ma, Journal of Statistical Physics, **26**, 221 (1981).

[9] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen, J Neurophysiol, **98**, 1064 (2007).

[10] A. Treves and S. Panzeri, Neural Computation, **7**, 399 (1995).

[11] J. Victor, Neural Computation, **12**, 2797 (2000).

[12] G. Miller, in *Information Theory in Psychology; Problems and Methods II-B*, edited by H. Quastler (Free Press, Glencoe, IL, 1955) pp. 95–100.

[13] P. Grassberger, Arxiv preprint physics/0307138 (2003).

[14] J. Victor, Physical Review E, **66**, 051903 (2002).

[15] S. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, Physical Review Letters, **80**, 197 (1998).

[16] A. Kraskov, H. Stögbauer, and P. Grassberger, Physical Review E, **69**, 066138 (2004).

[17] L. Kozachenko and N. Leonenko, Problems Information Transmission, **23**, 95 (1987).

[18] L. Paninski, Neural Computation, **15**, 1191 (2003).

[19] J. Bonachela, H. Hinrichsen, and M. Muñoz, Journal of Physics A: Mathematical and Theoretical, **41**, 202001 (2008).

[20] T. Schürmann, Journal of Physics A: Mathematical and General, **37**, L295 (2004).

[21] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, Physical Review E, **69**, 056111 (2004).

[22] I. Nemenman, F. Shafee, and W. Bialek, Arxiv preprint physics/0108025 (2001).

[23] D. H. Wolpert and D. R. Wolf, Phys Rev E, **52**, 6841 (1995).

[24] I. Nemenman, G. Lewen, W. Bialek, and R. van Steveninck, PLoS Computational Biology, **4** (2008).

[25] F. Montani, A. Kohn, M. A. Smith, and S. R. Schultz, Journal of Neuroscience, **27**, 2338 (2007).

[26] C. Shannon, Bell System Technical Journal, **27**, 379 (1948).

[27] S. Panzeri and A. Treves, Network: Computation in Neural Systems, **7**, 87 (1996).

[28] C. Magri, K. Whittingstall, V. Singh, N. Logothetis, and S. Panzeri, BMC neuroscience, **10**, 81 (2009).

[29] P. Grassberger, Physics Letters A, **128**, 369 (1988).

[30] I. Nemenman, Entropy, **13**, 2013 (2011).

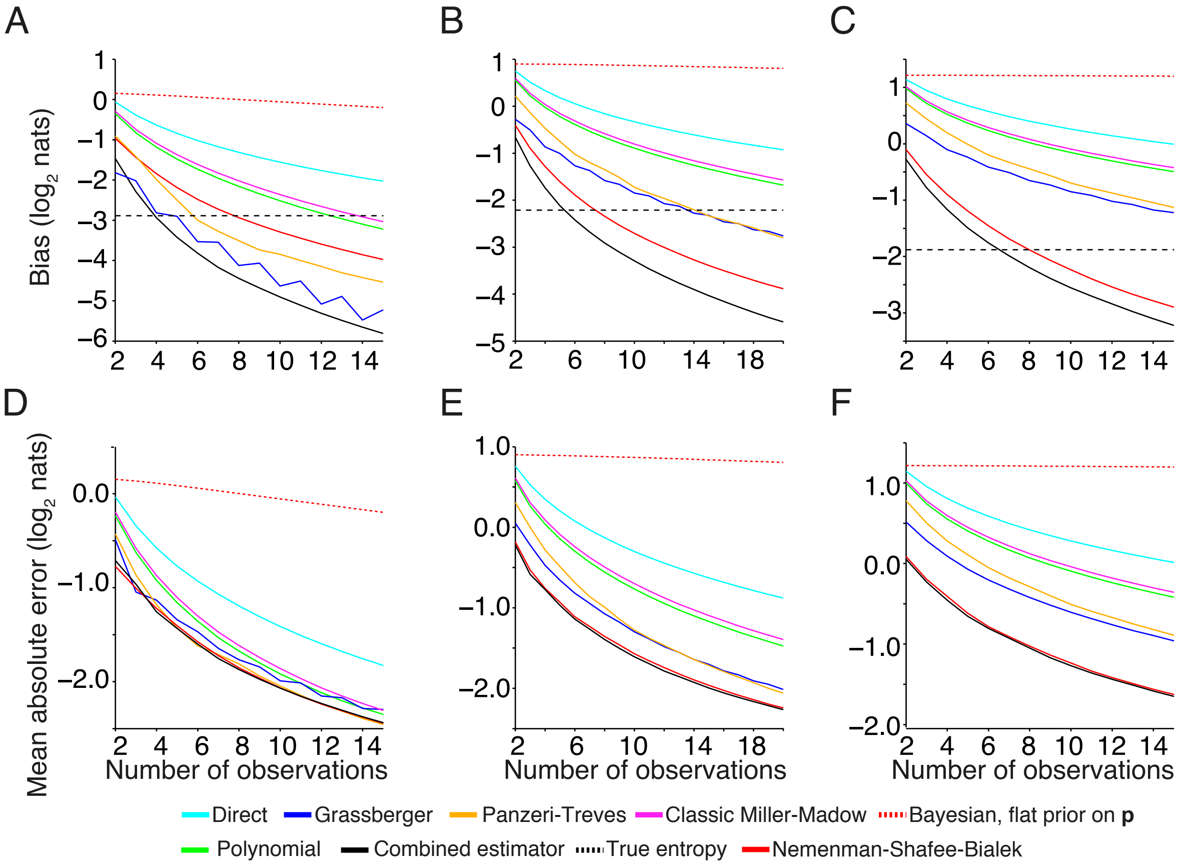

FIG. 1. Bias and average mean absolute error of various entropy estimators. (A) Average $\log_2$ transformed bias across the $[0, \log(M)]$ nats interval (*y-axis*), as a function of the number of observations n (*x-axis*), with the number of states $M = 16$. Black: our new combined $\hat{H}_{\mathrm{comb}}(\boldsymbol{n})$ entropy estimator (eq 27). Green: our new polynomial estimator $\hat{T}(\boldsymbol{n})$ (eq. 10). Red: NSB estimator [21]. Dashed red: Bayesian estimator with uniform prior on $\boldsymbol{p}$ [23]. Blue: Grassberger estimator [13]. Orange: Panzeri-Treves estimator [27]. Purple: classic Miller-Madow estimator [12]. Cyan: plugin estimator (eq. 1). Dashed black: 10% of the average entropy across the $[0, \ln(M)]$ nats interval. (B-C) As in (A), but now for $M = 81, 225$. (D) Average mean absolute error of the entropy estimates, averaged across the $[0, \log(M)]$ nats interval, with $M = 16$. Legends are similar to (A-C). (E-F): Similar to (D), but now for $M = 81, 225$.

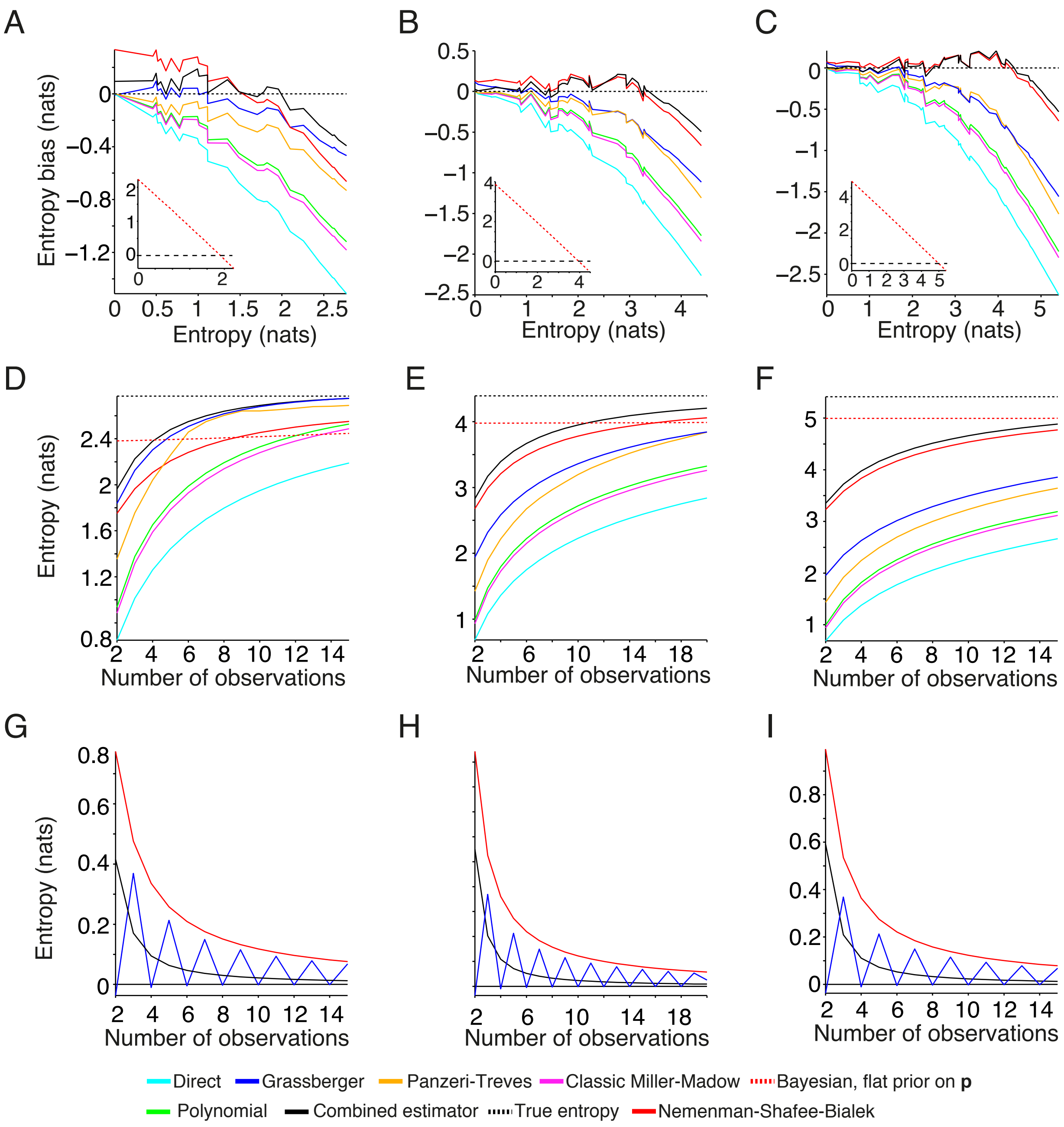

FIG. 2. (A-C) Entropy estimator bias for various levels of the entropy. Vectors of probabilities were generated by varying the concentration parameter β of a Dirichlet distribution with steps of 1/10 and drawing one vector of probabilities $\boldsymbol{p}$ from each Dirichlet distribution. The *x-axis* corresponds to the entropy of $\boldsymbol{p}$. The *y-axis* corresponds to the estimator bias in nats. Black: our new combined $\hat{H}_{\text{comb}}(\boldsymbol{n})$ entropy estimator (eq 27). Green: our new polynomial estimator $\hat{T}(\boldsymbol{n})$ (eq. 10). Red: NSB estimator [21]. Dashed red: Bayesian estimator with uniform prior on $\boldsymbol{p}$ [23]. Blue: Grassberger estimator [13]. Orange: Panzeri-Treves estimator [27]. Purple: classic Miller-Madow estimator [12]. Cyan: plugin estimator (eq. 1). Dashed black: to be estimated entropy of $\boldsymbol{p}$. (A-C) correspond to number of states $M = 16, 81, 225$ and number of observations $n = \sqrt{M} = 4, 9, 15$ respectively. (D-F) Expected value of various entropy estimators (*y-axis*) as a function of the number of observations n with the expected entropy of $\boldsymbol{p}$ equal to $\ln(M) - 1/1000$, i.e. close to the maximum entropy. (D-F) correspond to number of states $M = 16, 81, 225$, respectively. (G-I) Similar to (D-F), but now for the expected entropy of $\boldsymbol{p}$ equal to 1/10000, i.e. close to the minimum entropy.