2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

ON THE PERFORMANCE OF HISTOGRAM-BASED ENTROPY ESTIMATORS

Ciprian Doru Giurcăneanu[1,2]*, Panu Luosto*[3]*, Petri Kontkanen*[2,3]

[1]Department of Statistics, University of Auckland, New Zealand
[2]Helsinki Institute for Information Technology, HIIT, Finland
[3]Department of Computer Science, University of Helsinki, Finland
Email: c.giurcaneanu@auckland.ac.nz, {panu.luosto,petri.kontkanen}@cs.helsinki.fi

ABSTRACT

Histograms are widely used for estimating the density of a continuous signal from existing data. In some practical applications, they are also employed for entropy estimation. However, a histogram involves implicitly a discretization procedure because the unknown density is approximated by a piecewise constant density model. In the previous literature, the impact of the discretization procedure on the accuracy of the entropy estimate was either ignored or evaluated in the particular case of a regular histogram, in which all bins are equally wide. In this work, we provide bounds on the performance of the histogram-based entropy estimators without relying on the restrictive assumptions which have been used by other authors. The proof of our theoretical results is mainly based on concentration inequalities which have been already employed to analyze the performance of histograms as density estimators. After establishing the theoretical results, we illustrate them by numerical examples.

Index Terms— Histogram, entropy estimation, concentration inequalities, bias, variance

1. INTRODUCTION

Researchers working in signal processing are often confronted with the task of entropy estimation for continuous signals. The problem occurs in various research areas like independent component analysis [1], detection of abrupt changes and blind deconvolution [2], analysis of EEG signals [3].

One possible solution is to estimate firstly the unknown density and then to employ the obtained result for estimating the entropy. This two-stage procedure involves either kernels or histograms. In this work, we investigate the performance of the method which relies on histograms. Before presenting our contributions, we discuss briefly previous attempts for evaluating the performance of the two categories of methods. Since it is beyond the scope of this study to provide a complete list of works, we mention here only the references [4] and [5]. The first one is focused on the histogram-based approach, while the second one shows how some techniques which have been originally applied in the case of histograms can be adapted to analyze the kernel entropy estimators.

As our focus is on histograms, let us remark that their use involves implicitly a discretization procedure because the unknown density is approximated by a piecewise constant density model. So far, the impact of the discretization procedure on the accuracy of the entropy estimate was either ignored [4] or evaluated in the particular case of regular histograms for which all bins have equal width [3].

We propose a new method of analysis that does not make any assumption on the width of the bins and allows us to treat in a unitary manner both the regular and the irregular histograms. To this end, we resort to some techniques which are based on concentration inequalities. For a more detailed description of the techniques as well as for an exemplification of how they can be applied in density estimation, we refer to [6] and [7, Ch. 7]. In this paper, we will demonstrate their usefulness in finding bounds for bias and variance of the entropy estimate. In what concerns the bounds for variance, our methodology is not limited to the application of the inequalities from [6, 7], but we also investigate some other techniques like those from [8, 9].

The rest of the paper is organized as follows. The most important definitions are introduced in the next section. Then the bounds for the bias and the variance are given in Section 3 and Section 4, respectively. The theoretical results are illustrated by numerical examples in Section 5. Section 6 concludes the paper.

2. NOTATION AND DEFINITIONS

We outline below the main definitions which are similar to those previously introduced in [6], [7, Ch. 7], [10, Ch. 5].

Density estimation: Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with common law P on a measurable space $(\mathcal{Z}, \mathcal{T})$. Under the hypothesis that P admits a density s_* with respect to μ, or equivalently, $s_* = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, we aim to estimate s_* from the measurements $\xi_1, \ldots, \xi_n$. In our settings, μ is the Lebesgue measure on $\mathcal{Z}$.

Histogram: Assuming that $\mathcal{Z}$ is a compact interval of $\mathbb{R}$, we take $\Lambda_M = \bigcup_{j=1}^{D_M} \mathcal{I}_j$, where $\mathcal{I}_{j_1} \bigcap \mathcal{I}_{j_2} = \emptyset$ for $j_1 \neq j_2$. In other words, Λ_M is a partition of the interval $\mathcal{Z}$ into D_M pieces. Additionally, $\mu(\mathcal{I}) > 0$ for all $\mathcal{I} \in \Lambda_M$. Furthermore, we consider the linear vector space of piecewise constant functions with respect to Λ_M: $\tilde{M} = \left\{ s = \sum_{\mathcal{I} \in \Lambda_M} \beta_{\mathcal{I}} \mathbf{1}_{\mathcal{I}} \, : \, (\beta_{\mathcal{I}})_{\mathcal{I} \in \Lambda_M} \in \mathbb{R}^{D_M} \right\}$. The most important is the subset M of the functions in $\tilde{M}$ that are densities with respect to Λ_M: $M = \left\{ s \in \tilde{M} \, : \, s \geq 0, \int_{\mathcal{Z}} s \mathrm{d}\mu = 1 \right\}$. Remark that, for an arbitrary set $\mathcal{A}$, $\mathbf{1}_{\mathcal{A}}$ denotes its indicator function.

Maximum likelihood (ML) estimator: For a measurable function f on $\mathcal{Z}$, $P_n(f) = n^{-1} \sum_{i=1}^{n} f(\xi_i)$ is the empirical distribution associated to the samples $\xi_1, \ldots, \xi_n$. Given the model M, the ML estimator is that particular function $s \in M$ which minimizes $P_n(-\log s) = n^{-1} \sum_{i=1}^{n} [-\log s(\xi_i)]$. Observe that $\log(\cdot)$ stands for the natural logarithm. It can be easily proved (see [10, Ch. 5]) that the expression of the ML estimator is $\hat{s}_n(M) = \sum_{\mathcal{I} \in \Lambda_M} \frac{P_n(\mathcal{I})}{\mu(\mathcal{I})} \mathbf{1}_{\mathcal{I}} = \frac{1}{n} \sum_{\mathcal{I} \in \Lambda_M} \frac{1}{\mu(\mathcal{I})} \left[\sum_{\mathcal{I} \in \Lambda_M} \mathbf{1}_{\mathcal{I}}(\xi_i) \right] \mathbf{1}_{\mathcal{I}}$. For writing the equations more compactly, the notation $P_n(\mathcal{I})$ is used

instead of $P_n(\mathbf{1}_{\mathcal{I}})$.

Kullback-Leibler (KL) projection: If P and P_s are the probability distributions on $(\mathcal{Z},\mathcal{T})$ with respect to μ of the densities s_* and s, respectively, then the KL divergence has the expression: $D_{KL}(P,P_s)=\int_{\mathcal{Z}}\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_s}\right)\mathrm{d}P=\int_{\mathcal{Z}}s_*\log\frac{s_*}{s}\mathrm{d}\mu$. Like in [7], we use the notation $D_{KL}(s_*,s)$ instead of $D_{KL}(P,P_s)$. Remark that the integral within the definition is not finite whenever there is an interval $\mathcal{I}\subset\mathcal{Z}$ for which $s_*(\mathcal{I})>0$ and s is identically zero on $\mathcal{I}$. Among all densities s which belong to a given model M, there is one which minimizes $D_{KL}(s_*,s)$. This particular density is called "the KL projection of s_* onto M" and is denoted by s_M. It can be expressed as follows: $s_M=\sum_{\mathcal{I}\in\Lambda_M}\frac{1}{\mu(\mathcal{I})}\left[\int_{\mathcal{Z}}s_*\mathbf{1}_{\mathcal{I}}\mathrm{d}\mu\right]\mathbf{1}_{\mathcal{I}}$ [10]. With the convention that $P(\mathcal{I})=P(\mathbf{1}_{\mathcal{I}})=\int_{\mathcal{Z}}s_*\mathbf{1}_{\mathcal{I}}\mathrm{d}\mu$, we get $s_M=\sum_{\mathcal{I}\in\Lambda_M}\frac{P(\mathcal{I})}{\mu(\mathcal{I})}\mathbf{1}_{\mathcal{I}}$. Moreover, for a measurable function f on $\mathcal{Z}$, we have $P(f)=\mathbb{E}[f(\xi)]$, where ξ stands for a generic random variable of law P on $(\mathcal{Z},\mathcal{T})$.

Entropy estimation: Assume that s_* is estimated by using the histogram model M. Then, it is natural to estimate the entropy $H_*=-\int_{\mathcal{Z}}s_*\log s_*\mathrm{d}\mu$ by computing $\hat{H}_n(M)=-\int_{\mathcal{Z}}\hat{s}_n(M)\log\hat{s}_n(M)\mathrm{d}\mu$. In the next sections, we focus on finding bounds for the bias and the variance of $\hat{H}_n(M)$. In our calculations, it is helpful to employ the following definitions which are taken from [6]. If Λ_M is the partition which corresponds to M and η is an arbitrary number from the interval $]0,1[$, then

$$\Omega_\eta(M) = \left\{\frac{|P_n(\mathcal{I})-P(\mathcal{I})|}{P(\mathcal{I})}\le\eta,\ \forall\mathcal{I}\in\Lambda_M\right\}, \quad (1)$$

$$\chi^2_n(M) = \sum_{\mathcal{I}\in\Lambda_M}\frac{[P_n(\mathcal{I})-P(\mathcal{I})]^2}{P(\mathcal{I})} \quad (2)$$

$$= \int_{\mathcal{Z}}\frac{[\hat{s}_n(M)-s_M]^2}{s_M}\mathrm{d}\mu.$$

Additionally, the notation $\Omega^c_\eta(M)$ is used for the event $\{|P_n(\mathcal{I})-P(\mathcal{I})|/P(\mathcal{I})>\eta$ for some $\mathcal{I}\in\Lambda_M\}$.

After these preliminaries, we present next some results on the bias and the variance of $\hat{H}_n(M)$.

3. BOUNDS FOR BIAS OF THE ENTROPY ESTIMATE

Our derivations are based on the following set of assumptions:

- **(A1)** $\mathcal{Z}=[0,1]$.
- **(A2)** For some positive real number ρ, $s_*(z)\ge\rho$ for all $z\in\mathcal{Z}$, and $\int_{\mathcal{Z}}s_*(\log s_*)^2\mathrm{d}\mu<\infty$.
- **(A3)** For the model M, the cut-points of the partition Λ_M belong to the grid $\mathcal{G}=\{q/N_n\ :\ 0\le q\le N_n\}$, where N_n is a positive integer which satisfies the inequality

$$N_n\le\frac{n/(\log n)^2}{w}. \quad (3)$$

In the inequality above, the constant w is not smaller than one. It is important to emphasize that the model M can be either a regular or an irregular histogram. With respect to **(A1)**, let us observe that the necessity of imposing constraints for the support of s_* was already noticed in [3]. In fact, the regular grid from [3] covers only an interval of length 6σ, where σ is the standard deviation for the probability density function whose entropy is evaluated. Assumptions similar to **(A2)** and **(A3)** have been used in connection with the density estimation problem (see [7, Ch. 7] and the references therein for a more detailed discussion). In the beginning of our analysis, we prove the auxiliary result which is outlined below:

Lemma 3.1. *On the set $\Omega_\eta(M)$, we have:*

$$\frac{1-\eta}{2(1+\eta)^2}\chi^2_n(M)\le D_{KL}(\hat{s}_n(M),s_M)\le\frac{1+\eta}{2(1-\eta)^2}\chi^2_n(M). \quad (4)$$

Proof. It can be shown that the double inequality is true by applying Lemma 2.3 from [6]:

$$\begin{aligned}\frac{1}{2}\int_{\mathcal{Z}}\min\{p,q\}\left(\log\frac{p}{q}\right)^2\mathrm{d}\mu &\le D_{KL}(p,q) \quad (5)\\ &\le\frac{1}{2}\int_{\mathcal{Z}}\max\{p,q\}\left(\log\frac{p}{q}\right)^2\mathrm{d}\mu,\end{aligned}$$

where the densities p and q have the property that $\min\{p(z),q(z)\}>0$ for all $z\in\mathcal{Z}$. The key point is to observe in (5) that the upper and lower bounds are the same for both $D_{KL}(p,q)$ and $D_{KL}(q,p)$. With this observation, the inequalities in (4) can be obtained straightforwardly from [6, Eq. (2.13)]. Finally, we note that Lemma 2.3 from [6] is a particular case of Lemma 1 from [11]. □

Next we will show how Lemma 3.1 can be used for finding bounds for the bias of the entropy estimate. The strategy is to control the bias on the set $\Omega_\eta(M)$ and at the same time to verify how fast $\mathbb{P}[\Omega^c_\eta(M)]$ converges to zero when n increases.

Proposition 3.1. *Let $\delta_M=\inf_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I})$. If $\delta_M\in]0,1[$ and the assumptions* **(A1)**-**(A2)** *are satisfied, then the following inequalities hold:*

$$\begin{aligned}-\frac{1+\eta}{(1-\eta)^2}\frac{D_M-1}{2n}-A\cdot\left(\mathbb{P}[\Omega^c_\eta(M)]\right)^{1/2}&\\ \le\mathbb{E}\left[\left(\hat{H}_n(M)-H_\star-D_{KL}(s_*,s_M)\right)\mathbf{1}_{\Omega_\eta(M)}\right] \quad (6)&\\ \le-\frac{1-\eta}{(1+\eta)^2}\frac{D_M-1}{2n}+B\cdot\left(\mathbb{P}[\Omega^c_\eta(M)]\right)^{1/2},&\end{aligned}$$

where $A=-n^{-1/2}\log\min\{\rho,\delta_M\}$, $B=\frac{1-\eta}{(1+\eta)^2}(\rho\delta_M)^{-1}-n^{-1/2}\log\min\{\rho,\delta_M\}$ and

$$\mathbb{P}[\Omega^c_\eta(M)]\le\frac{2}{\delta_M}\exp\left(-\frac{n\rho\delta_M\eta^2}{2(1+\eta/3)}\right). \quad (7)$$

Proof. According to hypotheses, both $\hat{H}_n(M)$ and H_* are finite. The starting point is the identity:

$$\begin{aligned}\hat{H}_n(M)-H_*&\\ =\int_{\mathcal{Z}}s_*\log\frac{s_*}{s_M}\mathrm{d}\mu-\int_{\mathcal{Z}}(\hat{s}_n(M)-s_*)\log s_M\mathrm{d}\mu&\\ -\int_{\mathcal{Z}}\hat{s}_n(M)\log\frac{\hat{s}_n(M)}{s_M}\mathrm{d}\mu&\\ =D_{KL}(s_*,s_M)-D_{KL}(\hat{s}_n(M),s_M)-\nu_n(\log s_M), \quad (8)&\end{aligned}$$

where $\nu_n=P_n-P$ is the *centered empirical measure* [7, Def. 7.1]. For the identity above, it is important to observe that $s_M(z)\ge\rho>0$ for all $z\in\mathcal{Z}$ (see **(A2)**).

The use of Lemma 3.1 yields for $\mathbb{E}\left[-D_{KL}(\hat{s}_n(M),s_M)\mathbf{1}_{\Omega_\eta(M)}\right]$ bounds which depend on $-\mathbb{E}[\chi^2_n(M)\mathbf{1}_{\Omega_\eta(M)}]$. The fact that $\mathbb{E}[\chi^2_n(M)]=(D_M-1)/n$ leads to the identity $-\mathbb{E}[\chi^2_n(M)\mathbf{1}_{\Omega_\eta(M)}]=$

$-(D_M-1)/n+\mathbb{E}[\chi_n^2(M)\mathbf{1}_{\Omega_\eta^c(M)}]$, which implies that $-(D_M-1)/n \leq -\mathbb{E}[\chi_n^2(M)\mathbf{1}_{\Omega_\eta(M)}] \leq -(D_M-1)/n+2(\rho\delta_M)^{-1}\mathbb{P}[\Omega_\eta^c(M)]$. The upper bound for $-\mathbb{E}[\chi_n^2(M)\mathbf{1}_{\Omega_\eta(M)}]$ can be easily obtained by observing in (2) that

$$\begin{aligned}\chi_n^2(M) &= \sum_{\mathcal{I}\in\Lambda_M}\left[\frac{P_n^2(\mathcal{I})}{P(\mathcal{I})}-2P_n(\mathcal{I})+P(\mathcal{I})\right]\\ &\leq -1+\left(\sup_{\mathcal{I}\in\Lambda_M}\frac{1}{P(\mathcal{I})}\right)\sum_{\mathcal{I}\in\Lambda_M}P_n^2(\mathcal{I})\\ &\leq \sup_{\mathcal{I}\in\Lambda_M}\frac{2}{P(\mathcal{I})}\leq\frac{2}{\rho\delta_M}.\end{aligned}$$

We note in passing that the above inequality has been also employed in [6, p. 7].

It is clear that $\mathbb{E}[-\nu_n(\log s_M)\mathbf{1}_{\Omega_\eta(M)}]=\mathbb{E}[\nu_n(\log s_M)\mathbf{1}_{\Omega_\eta^c(M)}]$ because $\mathbb{E}[\nu_n(\log s_M)]=0$. Then we have the chain of inequalities:

$$\begin{aligned}&\left|\mathbb{E}[\nu_n(\log s_M)\mathbf{1}_{\Omega_\eta^c(M)}]\right|\\ &\leq \left(\mathbb{E}\left[(\nu_n(\log s_M))^2\right]\right)^{1/2}\left(\mathbb{P}[\Omega_\eta^c(M)]\right)^{1/2} \quad (9)\\ &\leq \left[\frac{1}{n}\int_{\mathcal{Z}}s_*(\log s_M)^2\mathrm{d}\mu\right]^{1/2}\left(\mathbb{P}[\Omega_\eta^c(M)]\right)^{1/2} \quad (10)\\ &\leq \frac{1}{\sqrt{n}}\log\left(\max\left\{\frac{1}{\rho},\frac{1}{\delta_M}\right\}\right)\left(\mathbb{P}[\Omega_\eta^c(M)]\right)^{1/2}. \quad (11)\end{aligned}$$

Remark that (9) is a straightforward consequence of the Cauchy-Schwarz inequality. More details on the proof of (10) can be found below. The inequality in (11) is based on the fact that, for all $\mathcal{I}\in\Lambda_M$, we have $\rho\leq s_M(\mathcal{I})=P(\mathcal{I})/\mu(\mathcal{I})\leq 1/\mu(\mathcal{I})\leq 1/\delta_M$.

To sketch the proof of the inequality in (10), we introduce the notation $t(z)=\log(s_M(z))$ for all $z\in\mathcal{Z}$. So,

$$\begin{aligned}\mathbb{E}\left[(\nu_n(t))^2\right] &= \mathbb{E}\left[(P_n(t))^2\right]-(P(t))^2\\ &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n t(\xi_i)\right)^2\right]-(P(t))^2\\ &= \frac{1}{n}P(t^2)-\frac{1}{n}(P(t))^2\\ &\leq \frac{1}{n}\int_{\mathcal{Z}}s_*t^2\mathrm{d}\mu.\end{aligned}$$

Collecting all the above inequalities, we obtain the result in (6). The inequality in (7) is based on Bernstein inequality [12], and it also appears in the previous literature (see, for example, [6, Eq. (2.9)]). □

Remark 3.1 In (6), the term $D_{KL}(s_*,s_M)$ takes only nonnegative values, and it becomes zero when s_* coincides with s_M. In fact, $D_{KL}(s_*,s_M)$ is that component of the bias which measures how well the estimated density is approximated by the histogram model M. The other component of the bias is mainly given by $-(D_M-1)/(2n)$ (see also the discussion below), and it depends on the sample size. This makes it to behave differently from $D_{KL}(s_*,s_M)$, which is independent of n. A similar decomposition of the bias has been presented in [3, Sec. 3] by using a different mathematical approach than the one employed to prove Proposition 3.1. The term $D_{KL}(s_*,s_M)$ plays a much more important role when M is not fixed, but is selected from a class of models. As this problem is not addressed here, we only mention that the interested reader can find some new results in [13].

Remark 3.2 According to an asymptotic result which is known since long time [14], the bias of $\hat{H}_n(M)-D_{KL}(s_*,s_M)$ is given by $-(D_M-1)/(2n)+o(1/n)$. A detailed discussion on various approaches which have been used to derive the asymptotic result can be found in [4, Sec. 4]. Here we focus on the relationship between this result and Proposition 3.1. The key point is to express $\mathbb{P}[\Omega_\eta^c(M)]$ as a function of n, and then to investigate its convergence rate. Under the hypothesis that the assumption **(A3)** holds, the inequality in (7) takes the particular form $\mathbb{P}[\Omega_\eta^c(M)]\leq \mathrm{UB}_n(w,\rho,c_\eta)$, where

$$\begin{aligned}\mathrm{UB}_n(w,\rho,c_\eta) &= \frac{2/w}{(\log n)^2}\left[n^{(w\rho c_\eta)\log n-1}\right]^{-1},\\ c_\eta &= \frac{\eta^2}{2(1+\eta/3)}.\end{aligned}$$

Since the product $w\rho c_\eta$ is strictly positive and independent of n, we have that

$$\lim_{n\to\infty}\frac{\mathrm{UB}_n(w,\rho,c_\eta)}{1/n}=0.$$

Hence, for a fixed D_M, the probability $\mathbb{P}[\Omega_\eta^c(M)]$ converges to zero faster than $(D_M-1)/(2n)$. However, a more careful analysis of the upper bound $\mathrm{UB}_n(w,\rho,c_\eta)$ should also take into account the influence of the parameters w, ρ and η. For instance, it is easy to prove that $c_\eta\in]0,3/8[$. The smaller is ρ, the slower is the decrease of $\mathrm{UB}_n(w,\rho,c_\eta)$ when $n\to\infty$. On the other hand, if one increases the value of w, then the maximum number of histogram bins will be reduced (see (3)) and $\mathrm{UB}_n(w,\rho,c_\eta)$ will lower more rapidly when n becomes larger.

The performance of the histogram-based entropy estimator is further investigated by finding upper bounds for the variance of $\hat{H}_n(M)$.

4. BOUNDS FOR VARIANCE OF THE ENTROPY ESTIMATE

In the demonstration of next result, we resort to some techniques that are different from those used in the previous section. More precisely, we apply Theorem 3 from [8] for proving the following proposition:

Proposition 4.1. *Under the assumptions* **(A1)**-**(A3)**, *we have:*

$$\mathrm{Var}[\hat{H}_n(M)]\leq\frac{9}{4}\frac{(\log n)^2}{n}. \quad (12)$$

Proof. Let us remark that

$$\begin{aligned}&-D_{KL}(\hat{s}_n(M),s_M)-\nu_n(\log s_M)\\ &= -\int_{\mathcal{Z}}\hat{s}_n(M)\log\frac{\hat{s}_n(M)}{s_M}\mathrm{d}\mu-\int_{\mathcal{Z}}[\hat{s}_n(M)-s_M]\log s_M\mathrm{d}\mu\\ &= -\sum_{\mathcal{I}\in\Lambda_M}P_n(\mathcal{I})\log\frac{P_n(\mathcal{I})}{P(\mathcal{I})}-\sum_{\mathcal{I}\in\Lambda_M}[P_n(\mathcal{I})-P(\mathcal{I})]\log\frac{P(\mathcal{I})}{\mu(\mathcal{I})}\\ &= \sum_{\mathcal{I}\in\Lambda_M}P(\mathcal{I})\log\frac{P(\mathcal{I})}{\mu(\mathcal{I})}-\sum_{\mathcal{I}\in\Lambda_M}P_n(\mathcal{I})\log\frac{P_n(\mathcal{I})}{\mu(\mathcal{I})}.\end{aligned}$$

By using the identity above together with (8), one can show without difficulties that the variance of $\hat{H}_n(M)$ equals the variance of the functional $\hat{F}_M(\xi_1,\ldots,\xi_n)=-\sum_{\mathcal{I}\in\Lambda_M}P_n(\mathcal{I})\log[P_n(\mathcal{I})/\mu(\mathcal{I})]$. It is important to notice how the value of the functional changes when the measurement ξ_i is replaced by $\xi_i'\neq\xi_i$ and all other

measurements are kept unchanged. Assuming that $\xi_i' \in \mathcal{I}_j$ and $\xi_i \in \mathcal{I}_k$, where both intervals $\mathcal{I}_j$ and $\mathcal{I}_k$ belong to Λ_M, furthermore we should consider two different cases:

(i)If $j = k$, we get

$$\begin{aligned}&\hat{F}_M(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)\\ &= \hat{F}_M(\xi_1, \ldots, \xi_{i-1}, \xi_i, \xi_{i+1}, \ldots, \xi_n).\end{aligned}$$

(ii) When $j \neq k$, the effect of the modification is twofold: $P_n(\mathcal{I}_j)$ increases from j/n to $(j+1)/n$, while $P_n(\mathcal{I}_k)$ decreases from k/n to $(k-1)/n$. Therefore, the functional is altered as follows:

$$\begin{aligned}&\Big|\hat{F}_M(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)\\ &\qquad - \hat{F}_M(\xi_1, \ldots, \xi_{i-1}, \xi_i, \xi_{i+1}, \ldots, \xi_n)\Big|\\ &= \Big| - \frac{j+1}{n}\log\frac{j+1}{n} + \frac{j}{n}\log\frac{j}{n} - \frac{k-1}{n}\log\frac{k-1}{n}\\ &\qquad + \frac{k}{n}\log\frac{k}{n} + \frac{1}{n}\log\frac{\mu(\mathcal{I}_j)}{\mu(\mathcal{I}_k)}\Big|\\ &\overset{(*)}{\leq} 2\frac{\log n}{n} + \frac{1}{n}\log\left(\frac{1}{\delta_M} - 1\right) \overset{(**)}{\leq} 3\frac{\log n}{n}.\end{aligned}$$

Note that in $(*)$ we have applied an inequality from [9, Sec. 2], and in $(**)$ we have used (3). Hence, we have shown that the modification of one sample point cannot alter the absolute value of the functional $\hat{F}_M(\cdot)$ by more than $c_n = (3\log n)/n$. According to [8, Th. 3], the result above leads to an upper bound for the variance of $\hat{F}_M(\cdot)$ which equals $(n/4)c_n^2$, and this concludes the proof. □

Remark 4.1 Note that Theorem 3 from [8] plays a key role in the proof of Proposition 4.1. The same theorem has been employed also previously in the context of entropy estimation (see, for example, [9]). However, the problem from [9] is slightly different from the one which we discuss here. More precisely, Antos and Kontoyiannis have assumed that the measurements take values on a countable alphabet and they are outcomes from a *discrete* distribution. Under these hypotheses, it has been demonstrated that the variance of the plug-in estimator for the entropy does not exceed $(logn)^2/n$. The bound which we have found in (12) is approximately double than the one from [9] since we have considered the impact of the discretization process. We cannot neglect this effect because the discretization itself is determined by the model M. However, in our proof we have assumed that the partition Λ_M contains only two intervals: one having length δ_M and the other one with length $1 - \delta_M$. Obviously, this is an extreme case and, for all other models M, the inequality (12) can be sharpen by replacing in $(*)$ the term $(1/n)\log(1/\delta_M - 1)$ with $\frac{1}{n}\log\frac{\max_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I})}{\min_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I})}$.

The natural question is if possible to obtain a sharper bound for the variance by using techniques like those from the proof of Proposition 3.1. In order to investigate this possibility, we bound the fourth order moment of $\chi_n(M)$. Unfortunately, we cannot apply straightforwardly the similar result which was proved in [7, p. 225], and this is why we show next all steps of the demonstration.

Lemma 4.1. *If the assumptions* **(A1)-(A3)** *are satisfied, then we have:*

$$\mathbb{E}\left[\chi_n(M)^4\right] \leq \frac{\kappa(\rho, \delta_M)}{(\log n)^4}, \tag{13}$$

where $\kappa(\rho, \delta_M)$ is a strictly positive constant and δ_M is the same as in Proposition 3.1.

Proof. We consider some more definitions from [6, 7]:

$$\begin{aligned}\varphi_{\mathcal{I}} &= P(\mathcal{I})^{-1/2}\mathbf{1}_{\mathcal{I}},\ \forall \mathcal{I} \in \Lambda_M,\\ \Phi_M &= \sum_{\mathcal{I}\in\Lambda_M}\varphi_{\mathcal{I}}^2,\\ V_M &= \sup_{\mathbf{a}\in\mathcal{A}_M}\left[\mathrm{Var}\left(\sum_{\mathcal{I}\in\Lambda_M} a_{\mathcal{I}}\varphi_{\mathcal{I}}(\xi_1)\right)\right],\end{aligned}$$

where $\mathcal{A}_M = \{\mathbf{a} \in \mathbb{R}^{|\Lambda_M|} : \sum_{\mathcal{I}\in\Lambda_M} a_{\mathcal{I}}^2 = 1\}$. It is easy to check that $\nu_n(\varphi_{\mathcal{I}}) = \int_{\mathcal{I}} \frac{\hat{s}_n(M) - s_*}{\sqrt{P(\mathcal{I})}}\mathrm{d}\mu = \frac{P_n(\mathcal{I}) - P(\mathcal{I})}{\sqrt{P(\mathcal{I})}}$, which leads to $\chi_n(M) = \left[\sum_{\mathcal{I}\in\Lambda_M}\nu_n^2(\varphi_I)\right]^{1/2}$ [6, p. 19].

More importantly, for any strictly positive constants ε and x, the following holds [7, p. 209]:

$$\begin{aligned}&\mathbb{P}\left[\chi_n(M) \geq (1+\varepsilon)\mathbb{E}[\chi_n(M)] + \sqrt{\frac{2V_M x}{n}} + c_\varepsilon\frac{\sqrt{||\Phi_M||_\infty}}{n}x\right]\\ &\quad \leq \exp(-x),\end{aligned} \tag{14}$$

where $c_\varepsilon = 2(\varepsilon^{-1} + 1/3)$. Now we focus on the relationship between D_M and the mathematical quantities involved in the left-hand side of the equation above. Firstly, the Jensen inequality leads to (see also [7, Eq. (7.10)]):

$$\mathbb{E}[\chi_n(M)] \leq \sqrt{\mathbb{E}[\chi_n^2(M)]} < \sqrt{D_M/n}. \tag{15}$$

Then, an upper bound for V_M can be readily obtained [7, p. 230]:

$$V_M \leq \sup_{\mathbf{a}\in\mathcal{A}_M}\mathbb{E}\left[\left(\sum_{\mathcal{I}\in\Lambda_M} a_{\mathcal{I}}\varphi_{\mathcal{I}}(\xi_1)\right)^2\right] \leq D_M. \tag{16}$$

From the hypotheses, we have that $\delta_M = \inf_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I}) > 0$ and $s_*(z) \geq \rho$ for all $z \in \mathcal{Z}$. So, $P(\mathcal{I}) \geq \rho\delta_M$ for all $\mathcal{I} \in \Lambda_M$, and

$$||\Phi_M||_\infty \leq D_M \sup_{\mathcal{I}\in\Lambda_M}\frac{1}{P(\mathcal{I})} \leq \frac{D_M}{\rho\delta_M}. \tag{17}$$

Finally, note that the assumption **(A3)** implies

$$\sqrt{\frac{D_M}{n}} \leq \frac{1}{\log n}. \tag{18}$$

Based on the results from (14)-(18), we conclude that there exits a constant $\kappa' > 0$ so as

$$\mathbb{P}\left[(\log n)\chi_n(M) > \kappa'(1+x)\right] \leq \exp(-x)\ \forall\, x > 0.$$

It is evident that κ' depends on both ρ and δ_M. From the equation above, we get the inequality in (13). □

Remark 4.2 To gain more insight on the inequality in (13), we resort to some well-known results. From (2), we have that $n\chi_n^2(M)$ is the same as the standard Chi-Square statistic [6, p. 5]. Moreover, when $n \to \infty$, the distribution of the standard Chi-Square statistic approaches $\chi^2(D_M - 1)$, where $D_M - 1$ is the number of degrees of freedom. Hence, for large n, $\mathbb{E}[\chi_n^4(M)] = (D_M^2 - 1)/n^2$ and, by making use of assumption **(A3)**, we get $\mathbb{E}[\chi_n^4(M)] \leq 1/(logn)^4$.

The inequality in (13) is further employed to bound the variance of $\hat{H}_n(M) - H_* - D_{KL}(s_*, s_M)$ on $\Omega_\eta(M)$.

Proposition 4.2. *If in addition to the hypotheses of Lemma 4.1 we have that $1/\rho \leq n$, then there exist two strictly positive constants $\kappa_1(\eta,\rho,\delta_M)$ and $\kappa_2(\eta,\rho,\delta_M)$ so as*

$$\begin{aligned}\mathrm{Var}\left[\left(\hat{H}_n(M) - H_* - D_{KL}(s_*,s_M)\right)\mathbf{1}_{\Omega_\eta(M)}\right] \\ \leq \frac{(\log n)^2}{n} + \frac{\kappa_1(\eta,\rho,\delta_M)}{\sqrt{n}\log n} + \frac{\kappa_2(\eta,\rho,\delta_M)}{(\log n)^4}.\end{aligned} \tag{19}$$

Proof. From the identity in (8), we readily obtain the following:

$$\begin{aligned}\mathrm{Var}\left[\left(\hat{H}_n(M) - H_* - D_{KL}(s_*,s_M)\right)\mathbf{1}_{\Omega_\eta(M)}\right] \\ \leq \mathbb{E}\left[\left(\nu_n(\log s_M) + D_{KL}(\hat{s}_n(M),s_M)\right)^2\mathbf{1}_{\Omega_\eta(M)}\right].\end{aligned}$$

By using the inequalities from the proof of Proposition 3.1, we get:

$$\begin{aligned}\mathbb{E}\left[\nu_n^2(\log s_M)\mathbf{1}_{\Omega_\eta(M)}\right] &\leq \mathbb{E}\left[\nu_n^2(\log s_M)\right] \\ &\leq \frac{\left[\log\max\{1/\rho,1/\delta_M\}\right]^2}{n} \leq \frac{(\log n)^2}{n}.\end{aligned}$$

Similarly, we have:

$$\begin{aligned}\left|\mathbb{E}\left[\nu_n(\log s_M)D_{KL}(\hat{s}_n(M),s_M)\mathbf{1}_{\Omega_\eta(M)}\right]\right| \\ \leq \frac{\log n}{\sqrt{n}}\left(\mathbb{E}\left[D_{KL}^2(\hat{s}_n(M),s_M)\mathbf{1}_{\Omega_\eta(M)}\right]\right)^{1/2}.\end{aligned}$$

Moreover, Lemma 3.1 implies that the following inequality holds on $\Omega_\eta(M)$: $D_{KL}^2(\hat{s}_n(M),s_M) \leq \frac{(1+\eta)^2}{4(1-\eta)^4}\chi_n^4(M)$. The application of Lemma 4.1 concludes the proof. □

Remark 4.3 As we already know from Remark 3.2, $\mathbb{P}(\Omega_\eta^c(M))$ goes very fast to zero when n increases. This recommends to compare the result in (19) with the one from (12). It is straightforward to verify numerically that, for the sample sizes currently available in practical situations, namely $n \in [50, 10^5]$, we have $\max\{1/(\sqrt{n}\log n),\ 1/(\log n)^4\} < (1/4)(\log n)^2/n$. However, we cannot conclude that (19) provides a better bound than (12) because the magnitudes of the constants $\kappa_1(\eta,\rho,\delta_M)$ and $\kappa_2(\eta,\rho,\delta_M)$ are unknown.

5. NUMERICAL EXAMPLES

The theoretical results established in the previous sections will be illustrated next by considering two different distributions. The first one is obtained by truncating the normal distribution with mean $1/2$ and variance $1/16$ to the interval $[0,1]$ and then re-normalizing it so as to integrate to unity. Hereafter, we will refer to this distribution as TN (truncated normal). The second distribution called TE is generated by applying the same techniques as in the previous case, with the major difference that this time we have started from an exponential distribution with mean $1/3$. Both TN and TE are plotted in Figures 1(a) and 1(d) together with their KL projections onto the considered models. Remark that the partition Λ_M for the histogram model of TN is given by the reunion of the intervals $[0,1/5]$, $]1/5,2/5]$, $]2/5,4/5]$ and $]4/5,1]$. It is evident that $\delta_M = 1/5$, and simple calculations show that $\rho \approx 0.226$. For TE, $\rho \approx 0.157$ and $\delta_M = 1/7$. The value of δ_M is a straightforward consequence of the fact that the intervals within the partition of the histogram model are $[0,2/7]$, $]2/7,3/7]$, $]3/7,4/7]$, $]4/7,6/7]$ and $]6/7,1]$.

As we already know from Remark 3.2, the upper bound for $\mathbb{P}[\Omega_\eta^c(M)]$ depends on the parameters w,ρ,η and the sample size n. Based on the discussion above, one can easily notice that $w = 1$ in our settings (see also (3)). As ρ is also known for both TN and TE, all that remains is to investigate the dependence of $\mathbb{P}[\Omega_\eta^c(M)]$ on η and n. To this end, we take $\{50,100,200,400,800\}$ to be the set of n-values for TN. Similarly, the set of sample sizes for TE is chosen to be $\{200,400,800,1600\}$. Furthermore, for each distribution and each sample size, 10^6 realizations are produced. The generated data are used to estimate empirically $\mathbb{P}[\Omega_\eta^c(M)]$ when η rises from 0.05 to 0.95. The results of this experiment are plotted in Figures 1(b) and 1(e). Observe that, for all n, the estimated values of $\mathbb{P}[\Omega_\eta^c(M)]$ are almost one when η is located in the vicinity of 0.05. However, the situation changes when η increases. To exemplify this trend, let us have a closer look at the case of TE when $\eta = 0.6$: $\mathbb{P}[\Omega_\eta^c(M)]$ is approximately 0.20 when $n = 200$, but it becomes as small as $7\cdot 10^{-5}$ when $n = 1600$. For the same value of η, in the plot drawn for TN, $\mathbb{P}[\Omega_\eta^c(M)]$ falls from 0.28 to less than 10^{-6} when n grows from 50 to 800.

Next we employ the values of $\mathbb{P}[\Omega_\eta^c(M)]$ which have been empirically computed for $\eta \in \{0.65, 0.70, 0.75\}$ to calculate the upper and lower bounds given in Proposition 3.1, and the results are shown in Figures 1(c) and 1(f). For the sake of comparison, we represent in the same plots, for each n, an approximation of $\mathbb{E}\left[\hat{H}_n(M)\right] - H_* - D_{KL}(s_*,s_M)$ which is obtained by replacing the expectation operator with an average over 10^6 realizations. Note that for all three selections of η that have been considered in this experiment, Proposition 3.1 does not provide tight bounds when $n = 50$ (for TN). The same is true when $n = 200$ (for TE). On the other hand, for both TN and TE, the difference between the upper and the lower bound diminishes very fast when n increases, and this behaviour is mainly caused by the rapid decrease of $\mathbb{P}[\Omega_\eta^c(M)]$ when n becomes larger (see Figures 1(b) and 1(e)).

Given that

$$\frac{\max_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I})}{\min_{\mathcal{I}\in\Lambda_M}\mu(\mathcal{I})} = 2$$

for the histogram model of TN as well as for the histogram model of TE, we make use of the result within Remark 4.1 to reduce the upper bound in (12) from $(3\log n)^2/(4n)$ to $(2\log n + \log 2)^2/(4n)$. Even after this improvement, the upper bound remains much larger than $\mathrm{Var}[\hat{H}_n(M)]$ computed from the same realizations that were considered in the experiments described above. For TN, the upper bound falls from 0.17 to 0.02 when n increases from 50 to 800, while the empirical estimate of $\mathrm{Var}[\hat{H}_n(M)]$ is smaller than 0.01 for all values of n. In the case of TE, the empirically calculated value of $\mathrm{Var}[\hat{H}_n(M)]$ does not exceed 0.005, whereas the upper bound varies from 0.055 to 0.01 as a function of the sample size.

6. CONCLUSION

In this paper, we have used concentration inequalities for analyzing the performance of histogram-based entropy estimators. The main advantage of our approach is that it takes into account the effect of the discretization process and, at the same time, allows to treat unitarily both the regular and the irregular histograms. An open problem is to improve the bounds provided in this work. Another possible line of research is to investigate how the results obtained in the 1D case can be extended to higher dimensions.

7. REFERENCES

[1] X.-L. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Process.*, vol. 58, pp. 5151–5164, 2010.

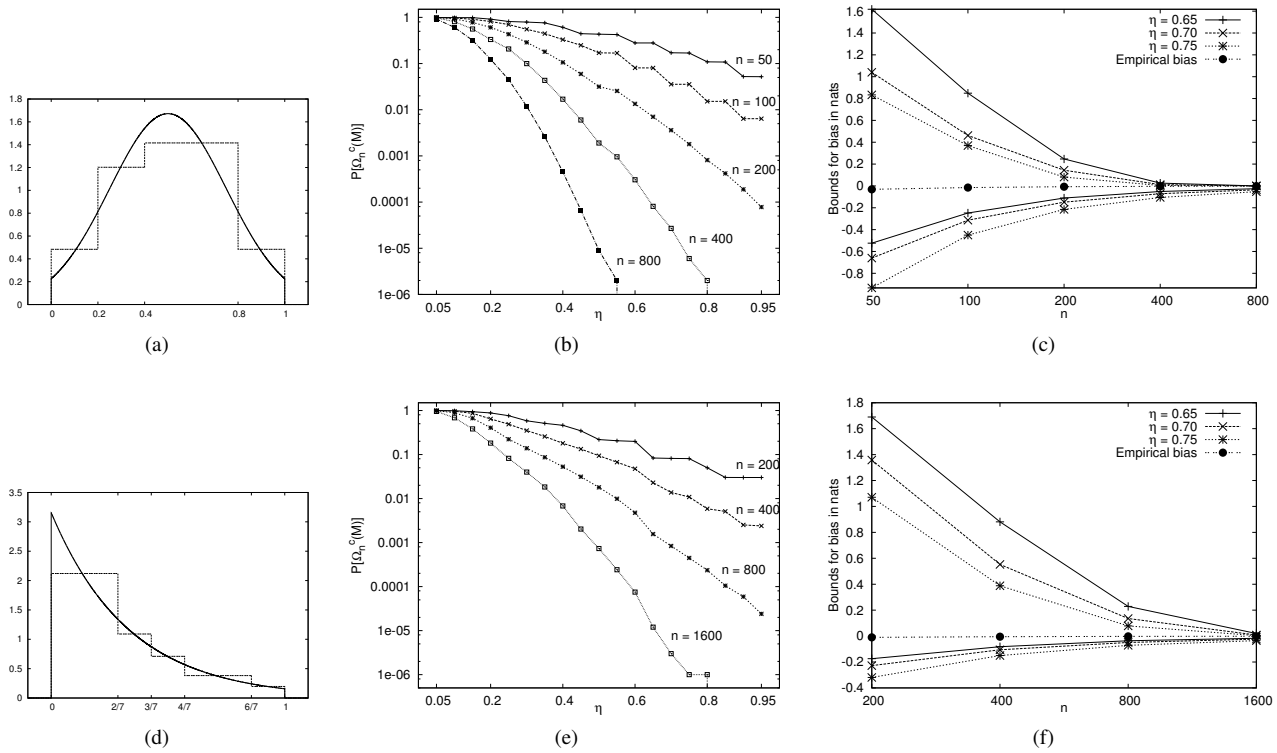

Fig. 1: Experimental results for two different distributions. All the plots within the first row are for the truncated normal (TN) distribution, while the second row is devoted to the truncated exponential (TE) distribution. In the first column, we show the two distributions together with their KL projections onto the considered models. We mention that $H_* = -0.1270$ and $D_{KL}(s_*, s_M) = 0.0267$ in the case of TN, while for TE we have $H_* = -0.3069$ and $D_{KL}(s_*, s_M) = 0.0237$. The graphs within the second column indicate the dependence of $\mathbb{P}[\Omega_\eta^c(M)]$ on η and n (see also definition in (1)). In the last column, we represent the upper and lower bounds for bias when the parameter η takes the values 0.65, 0.70 and 0.75 (see Proposition 3.1). In the plots within the third column, are also given the empirical values of $\mathbb{E}\left[\hat{H}_n(M)\right] - H_\star - D_{KL}(s_*, s_M)$ which are computed from 10^6 realizations for each sample size.

[2] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Trans. Signal Process.*, vol. 48, pp. 1687–1694, 2000.

[3] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Process.*, vol. 16, pp. 233–248, 1989.

[4] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1254, 2003.

[5] L. Paninski and M. Yajima, "Undersmoothed kernel entropy estimators," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4384 –4388, 2008.

[6] G. Castellan, "Modified Akaike's criterion for histogram density estimation," Technical Report #99.61, Univ. de Paris-Sud, France, 1999.

[7] P. Massart, *Concentration inequalities and model selection*, Springer Verlag, 2007.

[8] L. Devroye, "Exponential inequalities in nonparametric estimation," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed., pp. 31–44. Kluwer Academic Publishers, 1991.

[9] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Struct. Alg.*, vol. 19, pp. 163–193, 2001.

[10] A. Saumard, *Estimation par minimum de contraste régulier et heuristique de pente en sélection de modèles*, Ph.D. thesis, Univ. de Rennes 1, France, Oct. 2010.

[11] A. R. Barron and C.-H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Stat.*, vol. 19, pp. 1347–1369, 1991.

[12] D. Pollard, *Convergence of stochastic processes*, Springer Verlag, 1984.

[13] P. Luosto, CD Giurcăneanu, and P. Kontkanen, "Construction of irregular histograms by penalized maximum likelihood: a comparative study," in *Proc. IEEE Information Theory Workshop*, Lausanne, Switzerland, Sep. 2012 (to appear).

[14] G. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology II-B*, H. Quastler, Ed., pp. 95–100. Free Press, 1955.