ISIT'2013 1569719949

Some Worst-Case Bounds for Bayesian Estimators of Discrete Distributions

Steffen Schober
Institute of Communications Engineering
Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany
Email: steffen.schober@uni-ulm.de

***Abstract*—Let P be a distribution taking values in a finite set $\mathcal{X}$ with cardinality K. Given a sample $(X_1, X_2, \dots, X_n)$ a fundamental problem is to estimate the distribution P and its entropy $H(P)$. In practical applications often Bayesian estimators of P and $H(P)$ are used as they *may* give better results as for example the maximum likelihood estimator. The better performance can be achieved by choosing the *right* prior distribution on all possible distributions on $\mathcal{X}$. But choosing a wrong prior can be disastrous, especially for entropy estimation, as demonstrated by Nemenman, Shafee, and Bialek in 2001. In this work we give asymptotic worst-case results for Bayesian estimators using a symmetric Dirichlet prior. In particular, we show that estimators using the Laplace and Jeffrey prior can get arbitrarily close to P and H (in L_1 sense) for any distribution P if n scales with $K^{3/2+\delta}, \delta > 0$ as $n \to \infty$. For the Perks and Minimax prior this holds even if n scales with $K^{1+\delta}, \delta > 0$. As a *negative* result it is further shown that if the Laplace or the Jeffrey prior is used, there is always a distribution P such that the expected L_1 distance is bounded away from zero if n scales linear in K.**

I. Introduction

Let P be a distribution on the discrete set $\mathcal{X}, |\mathcal{X}| = K \in \mathbb{N}$ with probability mass function (pmf) $\{p(x), x \in \mathcal{X}\}$, where we allow that $p(x) = 0$ for some x. The entropy of P is defined as (see for example [1])

$$H(P) = -\sum_{x \in \mathcal{X}} p(x) \log p(x), \tag{1}$$

with $p \log p = 0$.

Given an i.i.d. sample $(X_1, X_2, \dots, X_n)$ drawn from P. The pmf of the maximum likelihood estimate $\hat{P}_{\text{ML}}$ (the empirical distribution) is obtained by

$$\hat{p}_{\text{ML}}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x), \tag{2}$$

where $I(\cdot)$ is the indicator of the event $\cdot$.

Bayesian estimators for the distribution P are obtained by assuming a *prior* distribution on the set of all possible probability distributions on $\mathcal{X}$. Throughout the paper we use a symmetric Dirichlet distribution $Dir(a, \cdots, a)$ on the probability simplex $(p_1, p_2, \dots, p_K), p_i \geq 0$ and $\sum p_i = 1$, with $\mathbb{R} \ni a > 0$ fixed as prior distribution. The estimated distribution $\hat{P}_B$ is obtained by the corresponding Bayesian estimate of the pmf of $\hat{P}_B$, given as

$$\hat{p}_{\text{B}}(x) = \frac{n\hat{p}_{\text{ML}}(x) + a}{n + Ka}. \tag{3}$$

From Eq. (2) and (3) entropy estimators are obtained by substituting the estimated $\hat{p}_{\text{B}}(x)$ into Eq. (1), for example,

$$\hat{H}_B(P) = H(\hat{P}_B) = -\sum_{x \in \mathcal{X}} \hat{p}_{\text{B}}(x) \log \hat{p}_{\text{B}}(x). \tag{4}$$

The estimators in Eq. (3) and Eq. (4) are often used with different choices for a. Some typical choices for a, which will be considered in the following, are shown in Table I.

a	Prior on P	Entropy estimator
0	no prior	Maximum Likelihood
$1/2$	Jeffrey [2]	Krichevsky & Trofimov [3]
1	Laplace [4]	Holste et al. [5]
$1/K$	Perks [6]	Schürman & Grassberger [7]
$\sqrt{n}/K$	minimax prior [8]	

TABLE I
TYPICAL PRIORS (SYMMETRIC DIRICHLET DISTRIBUTIONS WITH PARAMETER a) AND THE CORRESPONDING ENTROPY ESTIMATE.

Clearly, the *performance* of the Bayesian estimators crucially depend on the choice of a. For example, if we choose the correct prior, the Bayesian estimators have the lowest mean squared error among all estimators of P. But as shown for example in the case of entropy estimation, the estimators can dramatically fail if the wrong prior is chosen [9]. As stated in [9] "until the distribution is well sampled, our estimate of the entropy is dominated by the prior!" This leaves the open question, how many samples are enough in order to be *well sampled*.

In this paper this question is explored in more detail using an asymptotic approach similar to Paninski [10]. There it was shown that if n and K (more precisely K_n) tend to infinity such that ${}^{n}/_{K} > c > 0$ there is an entropy estimator such that the mean sqaured error tends to zero. In this paper we show that the Laplace and Jeffreys prior can get arbitrarily close to P and $H(P)$ (in L_1 sense) if n scales with $K^{3/2+\delta}$ where $\delta > 0$ as $n, K \to \infty$ assuming that the worst possible prior was chosen. For the Perks and Minimax prior this holds even for $n \geq K^{1+\delta}$. As a *negative* result it is further shown that for

the Laplace and Jeffrey prior that there is always a distribution P such that the expected L_1 distance is greater than zero if $n = \mathcal{O}(K)$. A similar result is implied for $|H(P) - \hat{H}_B(P)|$.

The rest of the paper is structured as follows. In Section II the main results are stated. The following section then presents the of the main results section.

II. Main Results

For P, Q being discrete probability distributions taking values in $\mathcal{X}$ (with $|\mathcal{X}| = K$) the L_q distance is defined as

$$L_q(Q, P) = \left(\sum_{x \in \mathcal{X}} |p(x) - q(x)|^q \right)^{1/q},$$

where we are mainly interested in $q = 1, 2$. An upper bound on the L_1 distance in terms of the L_2 distance is given by

$$L_1(P, Q) \leq \sqrt{K} L_2(P, Q). \tag{5}$$

The L_1 distance is related to the L_1 distances of entropies, namely if $L_1(P, Q) \leq 1/2$, then

$$|H(P) - H(Q)| \leq -L_1(P, Q) \log \frac{L_1(P, Q)}{K}, \tag{6}$$

e.g., [1]. Unfortunately, results for the L_1 distance are often difficult to obtain. But the following result is easily obtained by a similar result for the L_2 distance (which is shown in Section III-A) and by applying the upper bound Eq. (5):

Proposition 1. *Let*

$$\Xi(\epsilon) := (\epsilon - \sqrt{K}\mathbb{E}[L_2(\hat{P}_B, P)]) > 0.$$

It holds that

$$\Pr\left[L_1(\hat{P}_B, P) \geq \epsilon\right] \leq \exp\left(-\Xi(\epsilon)^2 \frac{(n+K)^2}{Kn}\right). \tag{7}$$

The proposition can be easily used to obtain bounds on the worst case error. The expected L_2 distances appearing above can be easily computed, upper bounds are summarized in Table II, exact values are given in Section III-B.

Proposition 1 is also used to obtain the following theorem.

Theorem 1. *Given an i.i.d. sample $(X_1, \ldots, X_n)$ where X_i is drawn according P which takes values in $\mathcal{X}$ with $|\mathcal{X}| = K$. Let δ, η be arbitrary positive constants and assume that n and K tend to infinity. If one of the following conditions is fulfilled*

1) *a is independent of K, n and $n \geq \eta \cdot K^{3/2+\delta}$*
2) *$a = 1/K$ or $a = \sqrt{n}/K$ and $n \geq \eta \cdot K^{1+\delta}$*

where a is the parameter specifying the pmf estimator $\hat{p}_B$ of $\hat{P}_B$ as given in Eq. (3). Then for any $\epsilon > 0$

$$\Pr\left[L_1(\hat{P}_B, P) \geq \epsilon\right] \to 0 \text{ as } n, K \to \infty. \tag{8}$$

Proof: We will prove the Theorem for the Laplace prior only, all other cases are similar. Let ϵ be fixed and assume w.l.o.g. that $n = K^{3/2+\delta}$ and $\delta \leq 1/2$. From Table II (see also (20)) we get in this case

$$\sqrt{K}\mathbb{E}[L_2(\hat{P}_B, Q)] \leq \frac{\sqrt{K^3}}{K+n}$$

Prior	Upper bound on $\mathbb{E}[L_2(\hat{P}_B, P)]$
Laplace	$(K+n)^{-1}\sqrt{\max(K^2, n)}$
Jeffrey	$(K+n)^{-1}\sqrt{\max(K^2/4, n)}$
Perks	$(n+1)^{-1}\sqrt{n}$
Minimax	$(\sqrt{n}+n)^{-1}\sqrt{n}$

TABLE II
Upper bounds on $\mathbb{E}[L_2(\hat{P}_B, P)]$ for the different priors.

monotonically decreases towards zero. Hence, there exists a n' such that $\epsilon \geq \sqrt{K}\mathbb{E}[L_2(\hat{P}_B, P)]$. This implies that $\Xi(\epsilon) > 0$ (in Eq. (7)) for all $n > n'$ and Eq. (7) gives the desired result. ∎

Theorem 1 has a counterpart for entropy estimation:

Theorem 2. *Under the same conditions as in Theorem 1 it holds that*

$$\Pr\left[|H(\hat{P}_B) - H(P)| \geq \epsilon\right] \to 0 \text{ as } n \to \infty. \tag{9}$$

Proof: Again, we prove the Theorem for the Laplace prior only where we assume that $n = K^{3/2+\delta}$ for $\delta > 0$. Let c be a real constant such that

$$0 < c < \min(1, 1/4 + \delta/2, \delta).$$

We start we the observation that if $L_1(\hat{P}_B, P) \leq \epsilon K^{-c} \leq 1/2$, then there is a K' such that for any $K \geq K'$ it holds that

$$|H(\hat{P}_B) - H(P)| \leq \epsilon.$$

To see this, substitute ϵK^{-c} into Eq. (6) which leads to

$$|H(\hat{P}_B) - H(P)| \leq \epsilon K^{-c}\left((1+c)\log K + \log \epsilon^{-1}\right),$$

which is smaller than ϵ for a large enough K. Therefore, to prove the result we want to apply Proposition 1 to bound $\Pr\left[L_1(\hat{P}_B, P) \geq \epsilon K^{-c}\right]$ for K being large enough. We observe that

$$\begin{aligned}
\Xi(\epsilon K^{-c}) &= K^{-c}(\epsilon - K^{1/2+c}\,\mathbb{E}[L_2(\hat{P}_B, P)]) \\
&\geq K^{-c}\left(\epsilon - K^{1/2+c}\frac{\sqrt{\max(K^2, n)}}{K+n}\right) \\
&= K^{-c}\left(\epsilon - \frac{K^{3/2+c}}{K + K^{3/2+\delta}}\right),
\end{aligned} \tag{10}$$

where the last line follows if we assume that $n = K^{3/2+\delta}$ (and say $\delta < 1/2$). If $c < \delta$, then (10) is always non-negative hence we can apply Proposition 1. But we should note that (10) tends to zero with K. In order to check that Proposition 1 gives the desired bound we need to check that exponent in Eq.(9) does not tend to zero as $K \to \infty$. The asymptotic behaviour of the negative exponent in (9) is

$$\begin{aligned}
\Xi(\epsilon K^{-c})^2 \frac{(n+K)^2}{Kn} &\sim \left(\epsilon - \frac{K^{3/2+c}}{K + K^{3/2+\delta}}\right)^2 \frac{K^{3+2\delta}}{K^{5/2+\delta+2c}} \\
&\sim K^{1/2+\delta-2c},
\end{aligned}$$

which tends to infinity if $c \leq 1/4 + \delta/2$ as required. ∎

Theorem 2 shows that the Bayesian entropy estimator using the Perks and Minimax prior come close to the estimator promised by [10]. One may wonder if it is possible to obtain a linear scaling. While we can not answer that question for the Perks and Minimax prior we will now show that at least for Laplace's and Jeffrey's prior this is not possible.

Proposition 2. *Let $a > 0$ be constant independent of n, K, and $K = cn$ for some constant $c > 0$. Then there is always a distribution P such that*

$$\mathbb{E}[L_1(\hat{P}_B, P)] > \Delta > 0,$$

as $n, K \to \infty$. Here Δ is a constant independent of n and K.

This implies a result for entropy estimation. In general $L_1(P, Q) > 0$ does not imply $|H(P) - H(Q)| > 0$, for example P and Q can be uniformly distributed on two disjoint subsets of $\mathcal{X}$. But the support of $\hat{P}_B$ and P has at least one common element, therefore, we obtain the following corollary:

Corollary 1. *Let $a > 0$ be constant independent of n, K, and $K = cn$ for some constant $c > 0$. Then there is always a distribution P such that*

$$\mathbb{E}[|H(\hat{P}_B) - H(P)|] > \Delta > 0$$

as $n, K \to \infty$. Here Δ is a constant independent of n and K.

III. Proofs

In this section we prove the open points from the previous section. Proposition 1 is shown in the following section, worst-case upper bounds for $\mathbb{E}[L_2(\hat{P}_B, P)]$ are given in III-B, finally the lower bounds for $\mathbb{E}[L_1(\hat{P}_B, P)]$ are given in III-C.

A. Proof of Proposition 1

Before we show Proposition 1 we will prove a closely related result:

Proposition 3. *Let*

$$\Theta(\epsilon) := (\epsilon - \mathbb{E}[L_2(\hat{P}_B, P)]) > 0.$$

Then

$$\Pr\left[L_2(\hat{P}_B, P) \geq \epsilon\right] \leq \exp\left(-\Theta(\epsilon)^2 \frac{(n+K)^2}{n}\right). \quad (11)$$

To prove the proposition we invoke the method of bounded differences:

Theorem 3 (Mc Diarmid [11])**.** *Let $X_1, \ldots, X_n$ be independent random variables, with X_k taking values in a set A_k for each k. Suppose that the function $f : \prod A_k \to \mathbb{R}$ satisfies*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_k$$

whenever the vectors $\mathbf{x}$ and $\mathbf{x}'$ differ only in the kth coordinate. Let Y be the random variables $f(X_1, \ldots, X_n)$. Then for any $t > 0$,

$$\Pr\left[Y - \mathbb{E}[Y] \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum c_k^2}\right).$$

Proof of Proposition 3: Let us consider the L_2 distance,

$$L_2(\hat{P}_B, P) = \left(\sum_{x \in \mathcal{X}} (\hat{p}_{\mathrm{B}}(x) - p(x))^2\right)^{1/2} \quad (12)$$

which is a random variable as $\hat{p}_{\mathrm{B}}(x)$ is a function of the i.i.d. sample $(X_1, \ldots, X_n)$. We start by the observation that we can write

$$\hat{p}_{\mathrm{B}}(x) = \frac{\lambda}{K} + (1 - \lambda)\hat{p}_{ML}(x), \quad (13)$$

with

$$\lambda = \frac{aK}{n + aK},\ 0 \leq \lambda \leq 1,$$

see e.g. [12]. Consider a (non-random) sample $(x_1, \ldots, x_n)$. Changing a single element x_k, say $x_k = i$ to $x_k = j$, will change $\hat{p}_{\mathrm{ML}}(i)$ and $\hat{p}_{\mathrm{ML}}(j)$ by not more than $1/n$ each. Hence, (12) will change by no more than $\sqrt{2}((1-\lambda)/n)$. Further,

$$\Pr\Big[L_2(\hat{P}_B, P) \geq \epsilon\Big] = \Pr\left[L_2(\hat{P}_B, P) - \mathbb{E}[L_2(\hat{P}_B, P)] \geq \Theta(\epsilon)\right].$$

By the assumptions of the proposition $\Theta(\epsilon)$ is non-negative hence invoking Theorem 3 with $\sqrt{2}((1-\lambda)/n) =: c_i$ proves the proposition. Note that the proof is closely related to the proof [13, Theorem 1]. ∎

Proof of Proposition 1: To show Proposition 1 we first note that for any $\delta > 0$

$$\Pr\left[L_2(\hat{P}_B, P) \geq \epsilon/\sqrt{K}\right] \leq \delta$$

implies that

$$\begin{aligned} &\Pr\left[L_2(\hat{P}_B, P) < \epsilon/\sqrt{K}\right] > 1 - \delta \\ \Rightarrow\ &\Pr\left[L_1(\hat{P}_B, P) < \epsilon\right] > 1 - \delta \\ \Rightarrow\ &\Pr\left[L_1(\hat{P}_B, P) \geq \epsilon\right] \leq \delta, \end{aligned}$$

where the second implication follows from Eq. (5) as the event $L_2 \leq \epsilon/\sqrt{K}$ implies $L_1 \leq \epsilon$. Hence, choosing $\epsilon = \epsilon'/\sqrt{K}$ in Proposition 3 directly implies Proposition 1. ∎

B. Worst-Case Upper Bounds on $\mathbb{E}[L_2]$

We will start with the following lemma:

Lemma 1.

$$\sum_x \mathbb{E}[\hat{p}_B(x)^2] = \frac{a^2K + n + 2an + (-1+n)nR(P)}{(aK+n)^2}.$$

(the expectation is with respect to the random sample) with

$$R(P) = \sum_{x \in \mathcal{X}} p(x)^2. \quad (14)$$

Proof: The lemma can be shown directly by summing up the square of the following equation

$$\hat{p}_{\mathrm{B}}(x) = \frac{\lambda}{K} + (1 - \lambda)\hat{P}_{ML}(x), \quad (15)$$

with $\lambda = \frac{aK}{n+aK}$, and using the first two moments of $\hat{p}_{\mathrm{ML}}$, i.e., $\mathbb{E}[\hat{p}_{\mathrm{ML}}(x)] = p(x)$ and

$$\mathbb{E}[\hat{p}_{\mathrm{ML}}(x)^2] = \frac{p(x) + p(x)^2(n-1)}{n}.$$

■

The following lemma gives the expected L_2 error:

Lemma 2.

$$\mathbb{E}[L_2(\hat{P}_B, P)] = \frac{\sqrt{n - a^2K + (a^2K^2 - n)R(P)}}{(aK+n)}. \quad (16)$$

Proof: In the following denote $L_2(\hat{P}_B, P) = \Delta$. From the definition

$$\mathbb{E}[\Delta^2] = \sum_x (\mathbb{E}[\hat{p}_{\mathrm{B}}(x)^2] - 2P(x)\mathbb{E}[\hat{p}_{\mathrm{B}}(x)] + P(x)^2).$$

The Lemma follows from Lemma 1 and

$$\mathbb{E}[\hat{p}_{\mathrm{B}}(x)] = (1-\lambda)p(x) + {}^{\lambda}/_{K}. \quad (17)$$

■

The quantity $R(P)$ appearing in the lemma above is closely related to the Renyi-Entropy of order 2, defined as $H_2(P) = -\log \sum_x p(x)^2$, see [14]. One can easily check, that it is minimal for the uniform distribution and maximal if $p(x) = 1$ for some x and zero else. Hence, the function is bounded by

$$\frac{1}{K} \leq R(P) \leq 1. \quad (18)$$

As a function of $R(P)$ the expected L_2 distance shows different behaviour depending on a:

$$\mathbb{E}[L_2(\hat{P}_B, P)] \text{ is } \begin{cases} \text{monotone increasing} & \text{if } a > \frac{\sqrt{n}}{K} \\ \text{independent of } R(P) & \text{if } a = \frac{\sqrt{n}}{K} \\ \text{monotone decreasing} & \text{if } a < \frac{\sqrt{n}}{K} \end{cases}. \quad (19)$$

1) Laplace and Jeffreys Prior: Here we study the priors for which a is a constant, so for example the Laplace or Jeffreys prior. The upper bound on the $\mathbb{E}[L_2(\hat{p}_{\mathrm{B}}, P)]$ distance follows from Eq. (16). From (19) we see that as long as $aK > \sqrt{n}$ the upper bound on the right hand side is monotone increasing in $R(P)$. Conversely if $aK < \sqrt{n}$ it is monotone decreasing with $R(P)$. Using the bounds on $R(P)$, Eq. (18), we obtain

$$\mathbb{E}[L_2(\hat{P}_B, P)] \leq \frac{1}{aK+n} \cdot \begin{cases} \sqrt{a^2K^2 - a^2K} & \text{if } aK > \sqrt{n} \\ \sqrt{n - a^2K} & \text{if } aK = \sqrt{n} \\ \sqrt{n - {}^{n}/_{K}} & \text{if } aK < \sqrt{n} \end{cases}$$

which is further upper bounded to yield

$$\mathbb{E}[L_2(\hat{P}_B, P)] \leq \frac{\sqrt{\max(a^2K^2, n)}}{aK+n}. \quad (20)$$

2) Minimax and Perks Prior: First let $a = {}^{\sqrt{n}}/_{k}$, again from (5) and (16) we obtain

$$\mathbb{E}[L_2(\hat{P}_B, P)] \leq \frac{\sqrt{n - {}^{n}/_{K}}}{(\sqrt{n}+n)} \leq \frac{\sqrt{n}}{\sqrt{n}+n}.$$

The behaviour of the Perks prior, i.e., $a = {}^{1}/_{K}$, can be expected to be similar as the minimax prior. Actually,

$$L_1(\hat{P}_B, P) \leq \sqrt{\frac{n - nR(P) + {}^{R(P)}/_{K} - {}^{1}/_{K^2}}{(n+1)^2}}.$$

As $a = \frac{1}{K} \leq \sqrt{n}/K$ the expected L_2 error is monotone decreasing with $R(P)$, therefore an upper bound is obtained for $R(P) = {}^{1}/_{K}$

$$L_1(\hat{P}_B, P) \leq \sqrt{\frac{n - {}^{n}/_{K}}{(n+1)^2}} \leq \sqrt{\frac{n}{(n+1)^2}}.$$

C. Lower bounds on the expected L_1 distance

To prove Proposition 2 we derive a lower bound on the expected L_1 distance in terms of the Hellinger distance. The latter is is defined as

$$H(P,Q) = \sqrt{\sum(\sqrt{p(x)} - \sqrt{q(x)})}.$$

We proceed with

$$\begin{aligned} \frac{1}{2}H(\hat{P}_B, P)^2 &= 1 - \sum_x \sqrt{\hat{p}_{\mathrm{B}}(x), p(x)} \\ &\geq 1 - \left(\sum \hat{p}_{\mathrm{B}}(x)^{\frac{p}{2}}\right)^{1/p} \left(\sum P(x)^{\frac{q}{2}}\right)^{1/q} \quad (21) \end{aligned}$$

for $1/p + 1/q = 1, 1 \leq p, q \leq \infty$, which follows from Hölder's inequality, i.e., for non-negative x_i, y_j

$$\sum_k x_k y_k \leq \left(\sum x_k^p\right)^{1/p} \left(\sum y_k^q\right)^{1/q}.$$

The expected squared Hellinger distance can be obtained from Eq. (21).

$$\begin{aligned} \frac{1}{2}\mathbb{E}[H(\hat{P}_B, P)^2] &\geq 1 - \left(\sum p(x)^{\frac{q}{2}}\right)^{1/q} \mathbb{E}[\left(\sum \hat{p}_{\mathrm{B}}(x)^{\frac{p}{2}}\right)^{1/p}] \\ &\geq 1 - \left(\sum p(x)^{\frac{q}{2}}\right)^{1/q} \left(\sum \mathbb{E}[\hat{p}_{\mathrm{B}}(x)^{\frac{p}{2}}]\right)^{1/p}, \end{aligned}$$

where the last line follows from Jensen's inequality as $(\cdot)^{1/p}$ is a concave function for $p > 1$ and the additivity of the expectation operator.

If we choose $p = 4$ (which implies $q = {}^{4}/_{3}$) we obtain

$$\frac{1}{2}\mathbb{E}[H(\hat{P}_B, P)^2] \geq 1 - \left(\sum p(x)^{\frac{2}{3}}\right)^{3/4} \left(\sum \mathbb{E}[\hat{p}_{\mathrm{B}}(x)^2]\right)^{1/4},$$

which leads to the lower bound on the expected L_1 distance

$$\mathbb{E}[L_1(\hat{P}_B, P)] \geq 2 - 2\left(\sum p(x)^{\frac{2}{3}}\right)^{3/4} \left(\sum \mathbb{E}[\hat{p}_{\mathrm{B}}(x)^2]\right)^{1/4}$$

as

$$H(P,Q)^2 \leq L_1(P,Q).$$

Let us now consider the distribution P' which is non-zero for only one element x, hence $p'(x) = 1$. In this case we have

$$\left(\sum p(x)^{\frac{2}{3}}\right)^{3/4} = 1$$

and from Lemma 1

$$\sum_x \mathbb{E}[\hat{p}_{\mathrm{B}}(x)^2] = \frac{a^2K + n + 2an + (-1+n)nR(P')}{(aK+n)^2}$$

with $R(P') = 1$. Suppose in this case that a is constant and that $n, K \to \infty$ such that $K = cn$ for some constant $c > 0$, then

$$\left(\frac{a^2K + n + 2an + (n-1)n}{(aK+n)^2}\right) \to \frac{1}{(1+ac)^2},$$

where the right hand side is positive and smaller one for any $a, c > 0$.

Hence, for any constant choice of a there exists at least one distribution, for which the expected L_1 distance is bounded away from zero if $K = cn$ as claimed.

ACKNOWLEDGEMENT

The author would like to thank V. Sidorenko for reading the manuscript.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., ser. Wiley Series in Telecommunications and Signal Processing. Wiley, Sept. 2006.

[2] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, Sept. 1946.

[3] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *Information Theory, IEEE Transactions on*, vol. 27, no. 2, pp. 199 – 207, Mar. 1981.

[4] P. Laplace, *Essai Philosophique Sur Les Probabilités*. H. Remy, 1829.

[5] D. Holste, I. Große, and H. Herzel, "Bayes' estimators of generalized entropies," *J. Phys. A*, vol. 31, pp. 2551–2566, 1998.

[6] W. Perks, "Some observations on inverse probability including a new indifference rule," *Journal of the Institute of Actuaries*, vol. 73, pp. 285–334, Jan. 1947.

[7] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos*, vol. 6, pp. 414–427, 1996.

[8] S. Trybula, "Some problems of simultaneous minimax estimation," *The Annals of Mathematical Statistics*, vol. 29, no. 1, pp. 245–253, Mar. 1958.

[9] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," in *Adv. Neural Inf. Proc. Syst*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, vol. 14.

[10] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *Information Theory, IEEE Transactions on*, vol. 50, no. 9, pp. 2200–2203, 2004.

[11] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, ser. London Mathematical Society Lecture Note Series, J. Siemons, Ed. Cambridge: Cambridge University Press, 1989, no. 141, pp. 148–188.

[12] A. Agresti and D. B. Hitchcock, "Bayesian inference for categorical data analysis," *Statistical Methods and Applications*, vol. 14, no. 3, pp. 297–330, Dec. 2005.

[13] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Struct. Algorithms*, vol. 19, no. 3-4, p. 163–193, Oct. 2001.

[14] A. Renyi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1960, pp. 547–561.