Estimating the Entropy of Discrete Distributions

András Antos[1]
Informatics Laboratory
Computer and Automation
Research Institute
of the Hung. Acad. of Sci.
H-1518 Lágymányosi u.11,
Budapest, Hungary.
email: antos@szit.bme.hu

Ioannis Kontoyiannis[2]
Division of Applied Mathematics
Brown University
Box F, 182 George Street
Providence, RI 02912, USA.
email: yiannis@dam.brown.edu
Tel: (401) 863-9246
Fax: (401) 863-1355.

Abstract — **Given an i.i.d. sample $(X_1,\ldots,X_n)$ drawn from an unknown discrete distribution P on a countably infinite set, we consider the problem of estimating the entropy of P. We show that the plug-in estimate is universally consistent and that, without further assumptions, *no* rate-of-convergence results can be obtained for *any* sequence of entropy estimates. Under additional conditions we get convergence rates for the plug-in estimate and for an estimate based on match-lengths. The behavior of the expected error of the plug-in estimate is shown to be in sharp contrast to the finite-alphabet case.**

I. Introduction

Suppose $P = \{p(i)\ ;\ i \in \mathcal{X}\}$ is an unknown discrete distribution on the countably infinite alphabet $\mathcal{X}$, and let $H = H(P)$ denote the entropy of P (in bits). Given an i.i.d. sample $(X_1,\ldots,X_n)$ drawn from P, we would like to be able to estimate H by some $H_n = H_n(X_1,\ldots,X_n)$, such that the error $|H_n - H|$ is typically small. We first ask whether universal estimates H_n exist (they do), and then we ask how fast they converge.

II. Consistency and Slow Rates

The *plug-in estimate* for H is defined by $\widehat{H}_n \triangleq H(p_n)$, where $p_n(i) = (1/n)\sum_{j=1}^n I_{\{X_j=i\}}$ is the empirical distribution induced by $(X_1,\ldots,X_n)$ on $\mathcal{X}$.

Proposition 1 *The plug-in estimate of H is strongly universally consistent, that is, $\widehat{H}_n \to H$ a.s. (as $n\to\infty$). For $H<\infty$, it is also consistent in L^2, that is, $\mathbf{E}\{(\widehat{H}_n - H)^2\}\to 0$, as $n\to\infty$.*

Theorem 1 *For any sequence $\{H_n\}$ of estimates for the entropy, and for any sequence $\{a_n\}$ of positive numbers converging to zero, there is a distribution P on $\mathcal{X}$ with $H(P)<\infty$, such that*

$$\limsup_{n\to\infty}\frac{\mathbf{E}\{|H_n - H|\}}{a_n} = \infty.$$

In [1], these two results are deduced from more general consistency and slow-rate results.

III. Convergence Rates

In view of Theorem 1, in order to obtain rate-of-convergence results, additional conditions need to be placed on the class of distributions P we consider.

Heuristics In the finite-alphabet case it is easy to see that $\mathbf{E}\{|\widehat{H}_n - H|\}$ decays like $1/\sqrt{n}$, and $\mathbf{E}\{(\widehat{H}_n - H)^2\}$ like $1/n$. We might expect that similar results should hold for an infinite alphabet $\mathcal{X}$, at least when $H^{(q)} = \mathbf{E}\{(-\log_2 p(X_1))^q\}$ is assumed to be finite for some $q\geq 2$. Theorem 2 shows that this is not at all the case. In fact, $\widehat{H}_n$ can tend to H at an arbitrarily slow algebraic rate even when $H^{(p)}<\infty$ for all p!

In our next result we restrict attention to distributions with tail probabilities decreasing like i^{-p} $(p>1)$. Without loss of generality we take $\mathcal{X} = \mathcal{N}$.

Theorem 2 *Assume that for some $p>1$ there exist positive constants $c_1,\ c_2>0$ such that $c_1/i^p \leq p(i)\leq c_2/i^p$, $i\in\mathcal{X}$. Then, for the plug-in estimate $\widehat{H}_n$ we have:*

$$\Omega\left(n^{-\frac{p-1}{p}}\right) = \mathbf{E}\{|\widehat{H}_n - H|\} \leq (\mathbf{E}\{(\widehat{H}_n - H)^2\})^{1/2} =$$

$$= \begin{cases} O\left(n^{-\frac{p-1}{p}}\right) & \text{if } p<2,\\ O\left(n^{-1/2}\log n\right) & \text{if } p\geq 2.\end{cases}$$

Given a sample $(x_1,x_2,\ldots,x_n)$ from $(X_1,\ldots,X_n)$, write $x_i^j = (x_i,x_{i+1},\ldots,x_j)$, $1\leq i\leq j\leq n$. For $n\geq 1$, define the match-lengths

$$L_n \triangleq \min\{1\leq L\leq n\ :\ x_1^L \neq x_{j+1}^{j+L}, \forall\, 1\leq j\leq n-L\},$$

and the corresponding entropy estimates

$$\widetilde{H}_n \triangleq \frac{\log_2 n}{L_n}.$$

Theorem 3

(a) *For $H<\infty$, we have $\widetilde{H}_n\to H$ a.s., as $n\to\infty$.*

(b) *If $H^{(2)}<\infty$, then*

$$\widetilde{H}_n = H + O\left(\frac{1}{\sqrt{\log n}}\right),$$

in probability.

(c) *If $H^{(2)}<\infty$ and $\mathbf{Var}\{\log_2 p(X_1)\}\neq 0$, then*

$$\mathbf{E}\{(\widetilde{H}_n - H)^2\}\geq[\mathbf{E}\{|\widetilde{H}_n - H|\}]^2 = \Omega\left(\frac{1}{\log n}\right).$$

(d) *If $H^{(4)} = \mathbf{E}\{(-\log_2 p(X))^4\}<\infty$, then*

$$\mathbf{E}\{(\widetilde{H}_n - H)^2\} = O\left(\frac{1}{\log n}\right).$$

References

[1] A. Antos, I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," Preprint, Sep. 2000.

[1] A.A. was supported in part by an Eötvös Scholarship.
[2] I.K. was supported in part by NSF grant #0073378-CCR.

0-7803-7123-2/01/$10.00 ©2001 IEEE