

Asymptotic Properties of Sample-Based Entropy, Information Divergence, and Related Metrics

Yuriy A. Reznik *

RealNetworks, Inc.

2601 Elliott Avenue, Seattle, WA 98121

E-mail: yreznik@ieee.org

Given a sample $w \in A^*$ produced by an unknown memoryless source S , it is common to estimate its entropy $H(S)$ by using [1]:

$$F(w) = - \sum_{\alpha \in A} \frac{r_{\alpha}(w)}{|w|} \log \frac{r_{\alpha}(w)}{|w|}, \quad (1)$$

where $r_{\alpha}(w)$ denotes the number of letters α in w , and $|w|$ denotes its total length.

In this abstract we show that a much higher precision (with $O(|w|^{-2})$ -vanishing error) can be achieved by adding a simple correction term:

$$F'(w) = F(w) + \frac{|A| - 1}{2|w|} \log e, \quad (2)$$

where $|A|$ is the cardinality of input alphabet. This follows immediately from:

Theorem 1. *The average sample-based entropy satisfies (with $|w| = n \rightarrow \infty$):*

$$\mathbf{E}\{F(w)\} = \sum_{w \in A^n} P_S(w) F(w) = H(S) - \frac{|A| - 1}{2n} \log e + \frac{\eta_1(S) - 1}{12n^2} \log e + O\left(\frac{1}{n^3}\right), \quad (3)$$

where $\eta_1(S)$ is a quantity depending on probabilities $P_S(\alpha)$ of the source S :

$$\eta_k(S) = \sum_{\alpha \in A} P_S^{-k}(\alpha) \quad (k = 1, 2, \dots). \quad (4)$$

Next, given two samples u and w from sources S and T correspondingly, we consider an estimate of the information divergence $D(S||T)$ between these sources:

$$G(u, w) = \sum_{\alpha \in A} \frac{r_{\alpha}(u)}{|u|} \log \frac{r_{\alpha}(u)/|u|}{r_{\alpha}(w)/|w|}, \quad (5)$$

defined under condition that $r_{\alpha}(w) = 0 \Rightarrow r_{\alpha}(u) = 0$. We report the following.

Theorem 2. *The average sample-based information divergence satisfies (asymptotically, with $|u| = n \rightarrow \infty$, and $|w| = t \rightarrow \infty$):*

$$\begin{aligned} \mathbf{E}\{G(u, w)\} &= \sum_{u \in A^n} \sum_{w \in A^t} P_S(u) P_T(w) G(u, w) \\ &= D(S||T) + \frac{|A| - 1}{2n} \log e + \frac{\eta_1(T) - 1}{2t} \log e - \frac{\eta_1(S) - 1}{12n^2} \log e \\ &\quad + \frac{1 - 6\eta_1(T) + 5\eta_2(T)}{12t^2} \log e + O\left(\frac{1}{n^3} + \frac{1}{t^3}\right). \end{aligned} \quad (6)$$

We also derive asymptotic expansions for sample-based entropy of mixtures and sample-based mutual information. These results are obtained by using technique discussed in [2].

References

- [1] R. E. Krichevsky, *Universal Compression and Retrieval* (Kluwer, Norwell, MA, 1993).
- [2] P. Flajolet, Singularity analysis and asymptotics of Bernoulli sums, *Theoretical Computer Science*, **215** (1999) 371-381.

*On leave from the Institute of Mathematical Machines and Systems, Kiev, Ukraine.