

Universal compression of unknown alphabets

Nikola Jevtić, Alon Orlitsky, Narayana Santhanam
ECE Department, UC San Diego, La Jolla, CA 92093
e-mail: {njevtic, alon, nsanthan}@ucsd.edu

Abstract — We consider universal compression of strings where the symbols are drawn independently according to the same unknown distribution over an unknown alphabet. We show that the order of the symbols can be conveyed using essentially as many bits as needed when the distribution is known in advance.

I. BACKGROUND

In many data-compression applications the source distribution is not known, except that it belongs to a given class \mathcal{P} of distributions, such as independent, identically distributed (iid), or Markov distributions.

The *worst-case redundancy* of \mathcal{P} is the lowest possible code-length increase due to not knowing the underlying probability, maximized over all strings and distributions in \mathcal{P} . Formally,

$$\hat{R}(\mathcal{P}, n) \stackrel{\text{def}}{=} \min_Q \max_{P \in \mathcal{P}} \max_{\bar{x} \in \mathcal{A}^n} \log \frac{P(\bar{x})}{Q(\bar{x})},$$

where \mathcal{A} is the underlying alphabet and Q is any probability distribution over \mathcal{A}^n .

Many results are known for distributions over finite alphabets. For example, for the collection \mathcal{P}_k of iid distributions over an alphabet of size k it has been shown [1] that

$$\lim_{n \rightarrow \infty} \left(\hat{R}(\mathcal{P}_k, n) - \frac{k-1}{2} \log \frac{n}{2\pi} - \log \frac{\Gamma(\frac{1}{2}k)}{\Gamma(\frac{k}{2})} \right) = 0. \quad (1)$$

In particular, the per symbol redundancy, $\hat{R}(\mathcal{P}_k, n)/n$, approaches zero as n increases.

II. UNKNOWN ALPHABETS

In many applications, such as text compression and facsimile transmission, the underlying alphabet is very large, often comparable to the length of the string, hence the limit in (1) does not apply. For the extreme case of infinite alphabets, Kieffer [2] showed that the per-symbol redundancy of any code is bounded away from zero.

We consider strings over unknown, possibly infinite, alphabets. Transmission of such strings can be viewed as consisting of two parts: a description of the symbols, and of the order in which they appear. For example, in the facsimile transmission of a document, the images corresponding to each character in the document are conveyed, then the order of the characters is transmitted.

We concentrate on the second task: communication of the order of the symbols. Formally, the *index* $I(x)$ of a symbol x in a string \bar{x} is one more than the number of distinct symbols preceding x in \bar{x} , and the *pattern* of \bar{x} is the string $I(x_1), \dots, I(x_n)$ of indices. For example, in the string “abracadabra”, $I(a) = 1$, $I(b) = 2$, $I(r) = 3$, $I(c) = 4$, and $I(d) = 5$, hence the pattern of “abracadabra” is 12314151231. Although not essential to our discussion, it can be shown that the number of patterns of length n is the n^{th} Bell Number.

Supported by NSF Grant 9815018.

III. REDUNDANCY

We consider the worst-case redundancy $\hat{R}_\rho(\mathcal{P}, n)$ of transmitting the pattern of a string generated according to a distribution in \mathcal{P} . We show that for the collection \mathcal{P}_u of iid distributions over an (unknown, possibly infinite) alphabet, the pattern can be transmitted essentially as efficiently as if the source distribution (and alphabet) were known in advance.

For $1 \leq i \leq n$, the i^{th} repetition number of an n -symbol string \bar{x} is the number $v_i(\bar{x})$ of symbols appearing i times in \bar{x} . It can be shown [4] that the probability distribution Q^* achieving $\hat{R}_\rho(\mathcal{P}_u, n)$ satisfies $Q^*(\bar{x}) \propto \prod_{i=1}^n \left(\frac{i}{n}\right)^{i v_i(\bar{x})}$. Hence

$$\hat{R}_\rho(\mathcal{P}_u, n) = \log \left(\sum_{\substack{v_1, \dots, v_n \geq 0 \\ \sum v_i = n}} \frac{n!}{n^n} \cdot \prod_{i=1}^n \frac{i^{i v_i}}{(i!)^{v_i} \cdot v_i!} \right).$$

IV. ASYMPTOTICS

For convenience define $\hat{r}_n = \exp(\hat{R}_\rho(\mathcal{P}_u, n))$. To calculate the asymptotic value of \hat{r}_n we evaluate the generating function

$$G(u) \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \frac{m^m}{m!} \hat{r}_m u^m,$$

where the first term is 1. For $|u| < \frac{1}{e}$,

$$G(u) = \exp \left(\sum_{m=1}^{\infty} \frac{m^m u^m}{m!} \right).$$

$G(u)$ satisfies the requirements of Hayman's Theorem [3], hence the asymptotic values of the coefficients of $G(u)$ can be evaluated, and the redundancies shown to be

$$\hat{r}_n = e^{\mathcal{O}(\sqrt{n})} \quad \text{and} \quad \hat{R}_\rho(\mathcal{P}_u, n) = \mathcal{O}(\sqrt{n}).$$

It follows that the indices of strings drawn according to iid distributions over unknown alphabets can be transmitted with zero asymptotic per-symbol worst-case redundancy. The same holds for average case redundancy.

V. ACKNOWLEDGEMENTS

We thank Yoav Freund for a question that helped initiate this work, and Suhas Diggavi, Krishnamurthy Viswanathan, and Junan Zhang for helpful discussions.

REFERENCES

- [1] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction”, *IEEE Trans. Inf. Theory*, 46:2, pp. 431-445, Mar. 2000.
- [2] J. C. Kieffer, “A unified approach to weak universal source coding”, *IEEE Trans. Inf. Theory*, 24:6, pp. 674-682, Nov. 1978.
- [3] W. K. Hayman, “A generalization of Stirling's Formula,” *J. reine angew. Math.*, No. 196, pp. 67-95, 1956.
- [4] Y. M. Shtarkov, “Universal sequential coding of single messages,” *Problems of Information Transmission*, vol. 23, No. 3, pp. 3-17, July-Sept. 1987.