The Twelvefold Way of Non-Sequential Lossless Compression

Taha Ameen ur Rahman, Alton S. Barbehenn, Xinan Chen, Hassan Dbouk, James A. Douglas, Yuncong Geng, Ian George, John B. Harvill, Sung Woo Jeon, Kartik K. Kansal, Kiwook Lee, Kelly A. Levick, Bochao Li, Ziyue Li, Yashaswini Murthy, Adarsh Muthuveeru-Subramaniam, S. Yagiz Olmez, Matthew J. Tomei, Tanya Veeravalli, Xuechao Wang, Eric A. Wayman, Fan Wu, Peng Xu, Shen Yan, Heling Zhang, Yibo Zhao, Yifan Zhang, Yibo Zhao, Sourya Basu, and Lav R. Varshney

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA
{sourya, varshney}@illinois.edu

Several frontiers of data science are producing huge amounts of data that are not simple sequences of distinguishable symbols, but have certain equivalence classes for patterns of symbols within which lossless representation does not require unique indices. As a simple example, the ordering of sequences of scientific data records are often irrelevant and so only the histogram or type class must be represented. Such representation problems arise for biological data, social graph structure, neural network architecture, and elsewhere due to the functional process generating or using the data. These non-sequential information sources may be studied implicitly via group theory and some may explicitly be characterized using interchange entropies.

Here we draw on enumerative combinatorics to explicitly find entropy bounds for a large class of possible kinds of non-sequential sources defined using functions between two finite sets. We specifically consider the twelvefold way in combinatorics due to Gian-Carlo Rota as a general classification of non-sequentiality that encompasses counting of permutations, combinations, multisets, and partitions, and therefore significantly generalizes from just irrelevance of order. The key mathematical step in information-theoretic analysis beyond known combinatorics is to characterize how probability distributions collapse under various invariances.

To demonstrate the approach, explicit entropy computations for all twelve settings are carried out for two basic sources: i.i.d. uniform and Bernoulli distributions. Comparisons among settings provide quantitative insight.

Going forward, it would be of interest to extend the approach developed in this paper to an even broader classification of group-theoretic invariances or of enumerative combinatorics. It is also of interest to develop data structures that perform appropriate “sorting” to allow more efficient encoding and decoding of the twelvefold way than just enumerative source coding.

An extended version of this work is available at https://arxiv.org/abs/2011.04069.