entropy 2021

sciforum

Abstract

On the CTW-Based Entropy Estimator †

Ronit Bustin

General Motors, Advanced Technical Center, Herzliya, Israel

† Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Published: 5 May 2021

Estimating the entropy of a sequence of discrete observations is a problem that raises in many different fields. There are numerous different applications of it, specifically in neuroscience, where entropy has been adopted as the main measure for describing the amount of information transmitted between neurons (see Gao et al., 2008 and reference therein). Gao et al., 2008 conducted a thorough comparison of the performance of the most popular and effective entropy estimators. They have shown that the context tree weighted (CTW) based estimator, which uses the probability estimation produced by the CTW lossless compression algorithm by Willems et al., 1995, repeatedly and consistently provides the most accurate results. The motivation for using the CTW probability for estimating the entropy is the well-known Shannon-McMillan-Breiman (SMB) result. However, the CTW probability is a result of a "twice universal" approach, meaning it is a weighted combination of the estimated probabilities of the sequence, over all possible bounded memory tree models (up to a predetermined maximum memory).

Motivated by this we examine the CTW based estimator from the view point of the CTW algorithm redundancy performance analysis (Willems et al., 1995). We define the SMB entropy as the normalized logarithm of the true probability, assuming a specific model for the source. We consider this finite length quantity to be the best possible estimator of the entropy given a specific model. By defining a random variable distributed over all possible bounded memory tree models we extend this definition and define the conditional SMB entropy. We bound the over estimation of the CTW-based estimator compared to both the conditional SMB entropy as well as the SMB entropy of a specific model. In both cases we show that the over estimation approaches zero according to O(log T/ T), where T is the length of the sequence.