Patterns in Random Words

Chaim Even-Zohar

Given a random text over a finite alphabet, we study the frequencies at which fixed-length words occur as subsequences. As the data size grows, the joint distribution of word counts exhibits a rich asymptotic structure. We investigate all linear combinations of subword statistics, and fully characterize their different orders of magnitude. Moreover, we establish the spectral decomposition of the space of word statistics of each order. We provide explicit formulas for the eigenvectors and eigenvalues of the covariance matrix of the multivariate distribution of these statistics. This framework includes as special cases several well-studied random variables from the combinatorial and statistical literature. Our techniques include algebraic tools such as representations and operators in the words algebra.
Joint work with Tsviqa Lakrec and Ran Tessler