Compression and entropy: A vignette on the work of Imre Simon

Volker Diekert
Universität Stuttgart

LATIN 2010, Oaxaca, Mexico, April 21, 2010

Imre made significant contributions to the theory of varieties of languages. Several of us remember the fascination of the late Samuel Eilenberg with Imre's new proof of the piecewise testable languages at the Spring School organized by the LITP in 1975.

A typical example is Imre's paper on the decidability of the finiteness of a semigroup of matrices over the tropical semiring.

At the heart of Imre's research is combinatorics of words. His characterization of the piecewise testable languages fits in this area, as well as his proof of Higman's number-theoretic theorem, and his investigation of the performance of the Ziv and Lempel compression algorithm whose explicit aim is to substitute a purely combinatorial approach for the more classical techniques of probability theory.

Imre writes each of his papers extremely carefully.

Personal relation with Imre

1. Words, Languages and Combinatorics, Kyoto, August 1990.
2. LATIN - conferences:
   - Valparaiso, Chile (1995), PC-member
   - Campinas, Brazil (1998), PC-member
   - Punta del Este, Uruguay (2000), PC-member
   - Cancun, Mexico (2002), PC-member
   - Buenos Aires, Argentina (2004), author
3. Ph.D. committee for Alair Pereira do Lago: Free Burnside semigroups, 1998.
4. Research stay in Stuttgart: June 1999.
5. Work on tight bounds for *Factorization Forests* (Manfred Kufleitner MFCS 2008)

Compression and Entropy

Invited Lecture at STACS 1992:
Georges Hansel, Dominique Perrin, and Imre Simon.
STACS 92, Lecture Notes in Computer Science **577** (1992)

Technical setting

Let $X \subseteq A^*$ be an infinite set of words to be compressed.
Let $\gamma : A^* \to \{ 0, 1 \}^*$ be a compressor (= injective function).

Worst-case compression rate:

$$\tau(X) = \limsup_{x \in X} \frac{|\gamma(x)|}{|x|}.$$

Average compression rate:

$$\tau_{\mathrm{ave}}(X) = \limsup_{n \to \infty} \frac{\sum_{x \in X \cap A^n} |\gamma(x)|}{n \cdot |X \cap A^n|}.$$

Topological entropy:

$$H(X) = \limsup_{n \to \infty} \frac{\log |X \cap A^n|}{n} \leq log\, |A|\,.$$

Main result of the 1992 paper by Hansel, Perrin, and Simon

$$H(X) \leq \tau_{\mathsf{ave}}(X) \leq \tau(X)$$

Let X be closed under factors, then the Ziv-Lempel-1978 compression ZL78 ($=$ compression using a dictionary arranged by prefixes) achieves the optimal bound:

$$H(X) = \tau_{\mathsf{ave}}^{\mathsf{ZL78}}(X) = \tau^{\mathsf{ZL78}}(X)$$

What about Ziv-Lempel-1977 compression ZL77 ?

Problem: For ZL77 two pointers must be stored.

It was conjectured that the results should hold for ZL77, too.

Diploma Thesis of Edgar Binder (Stuttgart 2008, unpublished):

Using $\log(n)$ bits for both pointers may indeed cause a factor 2 in the value of $\tau_{\mathsf{ave}}^{\mathsf{ZL77}}(X)$ (e.g. for $X = A^*$), but a Golomb-encoding of these pointers yields the desired optimality result for infinite factorial sets:

$$H(X) = \tau^{\mathsf{ZL77B}}(X)$$

Thank you

In memoriam: Imre Simon (August 14, 1943 – August 13, 2009)

You can help Wikipedia by expanding http://en.wikipedia.org/wiki/Imre_Simon