Faculty of Computer Science
Dalhousie University
6050 University Ave.
Halifax NS B3H 4R2
Canada

September 4, 2020

Dean's Office
Faculty of Mathematics and Computer Science
University of Wrocław
ul. F. Joliot-Curie 15
50-383 Wrocław
Poland

Re: evaluation of Michał Gańczorz's doctoral thesis

Dear Dean Jurdziński,

I have enjoyed reading Michał Gańczorz's thesis and found it to be an outstanding piece of work, clearly deserving of a doctorate. Mr Gańczorz has chosen four significant and thematically related papers, three of which have already appeared at important conferences, and written a clear exposition tying them together. I will make some comments on the next page but I leave it up to Mr Gańczorz whether or not to act on them, and I think his thesis can be accepted as it is.

In the first paper included in this thesis, Mr Gańczorz shows that a broad class of grammar-based compressors achieve good compression in terms of the kth-order empirical entropy of the input string, for reasonable values of k. Previous researchers had shown bounds in terms of empirical entropy but with larger coefficients, and in terms of the weaker property of universality; Mr Gańczorz's work has already inspired several significant follow-up papers by other authors. In the second paper, he shows a tight lower bound on the redundancy terms achievable with natural dictionary-based compression, when the compression is again measured in terms of the empirical entropy. The third paper shows how to parse an input string into phrases of bounded length such that the 0th-order empirical entropy of the parsing is minimized, considering the phrases as meta-characters and the the parsing as a string over those meta-characters. This result can be used to build data structures supporting fast random access to compressed strings, which is an important primitive for compressed data structures. Finally, the fourth paper shows how to compress labelled trees taking advantage of repetitive structure in both the shape of the tree and in the labels, in such a way that we can build efficient compact data structures on top of the compressed representations.

Mr Gańczorz's results have answered important open questions in data compression and compressed data structures, laid the foundations for long-term future research, and reinforced the University of Wrocław's reputation as a leading center for investigation into these subjects. Data compression is a well-established field, of course, but that makes such contributions all the more impressive. Compressed data structures were introduced relatively recently, with most techniques and applications having been introduced since the turn of the millennium, but they have already had a broad impact and the field is growing quickly, thanks to fundamental research such as that contained in this thesis.

Sincerely,

Travis Gagie, Dr. rer. nat.
Associate Professor
Dalhousie University

Travis Gagie

Digitally signed by Travis Gagie
DN: cn=Travis Gagie gn=Travis Gagie c=Canada l=CA o=Dalhousie University ou=Faculty of Computer Science e=travis.gagie@dal.ca
Reason: I am the author of this document
Location: Halifax, Canada
Date: 2020-10-17 19:20-08:00

Comments:

Sometimes there is a failure to clearly distinguish between entropy (a property of probability distributions) and empirical entropy (a property of strings). Similarly, it is implied that Huffman coding always works character-by-character, whereas it was first introduced to encode whole messages, given the probability distribution over the universe of possible messages. It would also be nice to have some discussion of, and comparison to, universality of codes in the sense that electrical engineers use it: they often assume a stationary and ergodic Markov source and show the compression rate approaches the entropy of the source with probability 1.

I found the following sentences on page 103 confusing:

> "This is of practical importance as the data structure, as the data structures with asymptotically smallest memory consumption, like [40], are very sophisticated, thus hard to implement and not always suitable for practical purposes. Thus we can choose theoretically inferior, but more practical data structure [2], we can even use a constant number of such data structures, as we are interested only in $O(|T'|)$ bound."

This seems inconsistent, as it starts off emphasizing practicality but then at the end ignores constant factors.

It would be nice if the author could add a couple of pages briefly discussing some recent developments: e.g.,

- `https://arxiv.org/abs/1910.02151` (Kociumaka et al.'s paper on measuring repetitiveness, appeared at LATIN 2020)
- `https://arxiv.org/abs/2006.01695` (Hucke et al.'s paper on empirical tree entropies, to appear at SPIRE 2020)
- `https://arxiv.org/abs/1910.07145` (our paper on random access to SLP-compressed texts, to appear at SPIRE 2020, which should have cited the author's STACS 2020 paper; my apologies)
- `https://arxiv.org/abs/2004.01120` (Prezza's post on indexing trees)

A brief discussion of locally-consistent parsing, prefix-free parsing and string-synchronizing sets would be interesting as well. Also, can Mr Gańczorz various analyses be extended to recompression?

For Chapter 3, is it possible to say anything about lower bounds that hold in the average case or with high probability? (Years ago I tried to investigate this for statistical compressors in `https://www.sciencedirect.com/science/article/pii/S157086671100075X`.) Also, block trees (see, e.g., `https://link.springer.com/chapter/10.1007/978-3-030-32686-9_31`) can support constant-time random access using $O(zn^\epsilon)$ space, where z is the number of phrases in the LZ77 parse, with the n^ϵ factor skirting Yu and Verbin's lower bound. It might be interesting to try to analyze them in terms of H_k, or try to adapt Yu and Verbin's lower bound for low H_k.

Does Theorem 1 on page 77 hold for all k simultaneously? On the same page, I do not see why $(|S|/\ell)\sum_{i=0}^{\ell-1} H_i(S) + O(\log|S|)$ is always less than $|S|H_k(S) + O((|S|/\log_\sigma|S|)(k\log\sigma + \log\log|S|))$; should there be some minimum for ℓ? For example, suppose $S = (01)^{n/2}$, so $H_0(S) = 1$ but $H_1(S) = 0$. Then for $\ell = 1$ and $k = 1$, we have $(|S|/\ell)\sum_{i=0}^{\ell-1} H_i(S) + O(\log|S|) = |S| + O(\log|S|)$ while $|S|H_k(S) + O((|S|/\log_\sigma|S|)(k\log\sigma + \log\log|S|)) = O(|S|\log\log|S|/\log|S|)$.

On page 18, I don't see why $|L| = o(|S|)$ implies $|L|H_0(L) = o(|S|)$ (and "$o(|S|)$" is written "$o(S)$").

The thesis is very well written but there are a few typos, such as the following:

- in the abstract, "work good" should be "work well";
- in 1.3.1, "their encodings' costs are close to", and "which output are close to entropy";
- in 1.3.3, "which we believe are of independent interest";
- in 1.5, "most of these structures" and "We probably cannot expect optimal solutions for either of these problems";
- on page 22, "any of the:" should be "any of the following:";
- in the abstract on page 56, "the those";
- in the abstract on page 76, there is a missing bracket in the penultimate line.