Weekplan: Compression

Inge Li Gørtz

References and Reading

[1] Introduction to Data Compression, Guy E. Blelloch.

[2] Improved Approximate String Matching and Regular Expression Matching on Ziv-Lempel Compressed Texts, Philip Bille, Rolf Fagerberg, and Inge Li Gørtz. In ACM Transactions on Algorithms, volume 6(1), pages 1-14, 2009.

[3] A Universal Algorithm for Sequential Data Compression, Jacob Ziv and Abraham Lempel: IEEE Transactions on Information Theory 23 (3): 337?34.

[4] Compression of Individual Sequences via Variable-Rate Coding, Jacob Ziv and Abraham Lempel. IEEE Transactions on Information Theory 24 (5): 530?536.

We recommend reading [1] section 5 and [2] section 2 in detail before the lecture. [3] and [4] provides background on the the Lempel-Ziv compression schemes.

Exercises

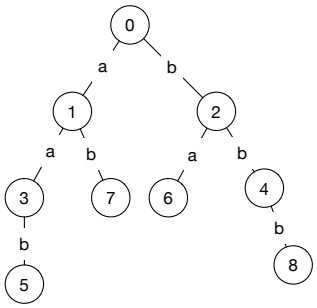
1 [w] **LZ77 and LZ78** Encode the string `bcabccabccab` using LZ77 and LZ78.

2 [w] **LZ77 decoding** Decode the following LZ77 factoring:

$$(0,0,a),(0,0,b),(2,1,a),(3,1,b),(4,8,b).$$

How does the LZ77 encoding look if self-referencing is not allowed?

3 [w] **LZ78 decoding** Decode the string represented by the trie below.



4 Compression rates Suppose you are given a text of size N. What is the best possible compressed file sizes achievable by LZ77 and LZ78.

5 Random access in LZ78 Consider a string compressed with LZ78. Explain how to report the character at a given position, fx. position 7, by looking at the trie. What extra information do you need?

6 Optimality of greedy parsing LZ77 always finds the longest substring we have seen before when constructing a new phrases. This is called greedy parsing. A natural question is wether it would be possible to encode the string with a smaller number of phrases if we did not have to take the longest match each time. Show that the greedy parsing in LZ77 is optimal in the sense that it always finds the smallest number of phrases.

7 Properties of LZ77 Given a string T, let $L_{77}(T)$ be the number of phrases of the LZ77 parsing of T. Show that the following properties are true for any characters a:

1. $L_{77}(TT) = L_{77}(T) + 1$;
2. $L_{77}(TT') \leq L_{77}(TaT') + 1$.
3. $L_{77}(TaT') \leq L_{77}(TT') + 1$.

Do the properties also hold for LZ78?

8 Compressed k-mismatch pattern matching In the k-mismatch problem we are given a text T of length N, a pattern P of length m, and an integer k, and the problem is to find all occurrences of the pattern in the text, each with at most k mismatches.

8.1 Explain how to solve the problem in $O(Nmk)$ using a data structure for longest common extensions.

8.2 Give an algorithm to solve the problem when the text compressed with LZ78.

9 LZ77 with sliding window Describe how to use suffix trees to implement LZ77 with sliding window on a string S of length n, with window size W. Note that LZ77 can only report a match that starts within the window.