Space Lower Bounds for Online Pattern Matching

Raphaël Clifford[1], Markus Jalsenius[1], Ely Porat[2], and Benjamin Sach[1]

[1] University of Bristol, Dept. of Computer Science, Bristol, UK
[2] Bar-Ilan University, Dept. of Computer Science, Ramat-Gan, Israel

Abstract. We present space lower bounds for online pattern matching under a number of different distance measures. Given a pattern of length m and a text that arrives one character at a time, the online pattern matching problem is to report the distance between the pattern and a sliding window of the text as soon as the new character arrives. We require that the correct answer is given at each position with constant probability. We give $\Omega(m)$ bit space lower bounds for L_1, L_2, L_∞, Hamming, edit and swap distances as well as for any algorithm that computes the cross-correlation/convolution. We then show a dichotomy between distance functions that have wildcard-like properties and those that do not. In the former case which includes, as an example, pattern matching with character classes, we give $\Omega(m)$ bit space lower bounds. For other distance functions, we show that there exist space bounds of $\Omega(\log m)$ and $O(\log^2 m)$ bits. Finally we discuss space lower bounds for non-binary inputs and show how in some cases they can be improved.

1 Introduction

We combine existing results with new observations to present an overview of space lower bounds for online pattern matching. Given a pattern that is provided in advance and a text that arrives one character at a time, the online pattern matching problem is to report the distance between the pattern and a sliding window of the text as soon as the new character arrives. In this formulation, the pattern is processed before the first text character arrives and once processed, the pattern is no longer available to the algorithm unless a copy is explicitly made.

This problem has recently gained a great deal of interest with breakthrough results given for exact matching and pattern matching under bounded Hamming distance (k-mismatch) [13]. For both problems it was shown that space sublinear in the size of the pattern is sufficient to give the correct answer at every alignment with high probability. These remarkable results immediately raise a number of significant unresolved questions. The first is for which other distance measures between strings might sublinear space randomised online algorithms be achievable and it is this question which we address here.

Our presentation is divided between what we term local and non-local online pattern matching problems. In the former case the distance function between

R. Giancarlo and G. Manzini (Eds.): CPM 2011, LNCS 6661, pp. 184–196, 2011.

a pattern P of length m and an m-length substring of the text T, starting at position i, is defined by

$$\textsc{LocalPM}_{(\oplus,\Delta)}(P,T) = \bigoplus_{j=0}^{m-1} \Delta(P[j], T[i+j])\,,$$

where $\oplus$ and Δ are both binary operators. In Section 4 we show $\Omega(m)$ bit space lower bounds for online pattern matching for the local problems of L_1, L_2, and Hamming distance as well as for any algorithm that computes the cross-correlation/convolution.

We then go on to show in Section 5 a space dichotomy for local online pattern matching problems of the form $d(i) = \bigwedge_{j=0}^{m-1} \Delta(P[j], T[i+j])$ where the range of Δ is $\{\textsc{True}, \textsc{False}\}$. Where the distance function Δ has wildcard-like properties (qv. Section 5), we give an $\Omega(m)$ space lower bound. Where it does not, we have $\Omega(\log m)$ and $O(\log^2 m)$ space bounds. This implies, for example, that online pattern matching with character classes [8] requires linear space.

In Section 6 we go on to consider all eight possible binary Boolean associative operators and give a complete classification in terms of their known upper and lower space bounds. One consequence is that determining if there is an exact "non-match", where the Hamming distance is the same as the pattern length, requires linear space in our online model. This bound also holds if, for example, only the parity of the Hamming distance is required. In Section 7 we then show how our techniques can be used to give linear space lower bounds for L_∞ online pattern matching. In Section 8 we discuss a possible approach to space lower bounds for inputs with large alphabets, focussing on the Hamming distance problem. Finally, in Section 9 we explore non-local problems and show $\Omega(m)$ bit space lower bounds for both online edit and swap distance.

2 Preliminaries and Related Work

Let Σ_P and Σ_T denote the pattern and text alphabet, respectively. We say that $\textsc{LocalPM}_{(\oplus,\Delta)}$ is *text independent* with respect to the pattern P if the value of $\textsc{LocalPM}_{(\oplus,\Delta)}$ is a constant independent of T. We say that $\textsc{LocalPM}_{(\oplus,\Delta)}$ is *pattern independent* with respect to a pattern P if there is a function Δ' such that $\Delta(x,y) = \Delta'(y)$ for all $(x,y) \in P \times \Sigma_\mathrm{T}$.

Example 1. Let $\Sigma_\mathrm{P} = \{x,y,z\}$, $\Sigma_\mathrm{T} = \{a,b,c\}$, $\oplus$ be the Boolean AND-operator and Δ be defined according to the table in Fig. 1, where 1 is $\textsc{True}$ and 0 is $\textsc{False}$. We can see that $\textsc{LocalPM}_{(\wedge,\Delta)}$ is text independent with respect to the pattern $P = xxyyxzxx$ as it always outputs 0. It is also pattern independent with respect to $P = yyzyyzzy$ as $\Delta(y,\alpha) = \Delta(z,\alpha)$ for all $\alpha \in \Sigma_\mathrm{T}$. In fact, for this particular definition of Δ, $\textsc{LocalPM}_{(\wedge,\Delta)}$ is either text or pattern independent with respect to any pattern P.

Suppose that $\textsc{LocalPM}_{(\oplus,\Delta)}$ is text independent with respect to a pattern P. Then any algorithm for $\textsc{LocalPM}_{(\oplus,\Delta)}$ on P requires at most $O(1)$ space after