




# The Dynamic $k$ -Mismatch Problem

Raphaël Clifford  

Department of Computer Science, University of Bristol, UK

Paweł Gawrychowski  

Institute of Computer Science, University of Wrocław, Poland

Tomasz Kociumaka  

University of California, Berkeley, CA, USA

Daniel P. Martin  

The Alan Turing Institute, British Library, London, UK

Przemysław Uznański  

Institute of Computer Science, University of Wrocław, Poland

---

## Abstract

The text-to-pattern Hamming distances problem asks to compute the Hamming distances between a given pattern of length  $m$  and all length- $m$  substrings of a given text of length  $n \geq m$ . We focus on the well-studied  $k$ -mismatch version of the problem, where a distance needs to be returned only if it does not exceed a threshold  $k$ . Moreover, we assume  $n \leq 2m$  (in general, one can partition the text into overlapping blocks). In this work, we develop data structures for the dynamic version of the  $k$ -mismatch problem supporting two operations: An update performs a single-letter substitution in the pattern or the text, whereas a query, given an index  $i$ , returns the Hamming distance between the pattern and the text substring starting at position  $i$ , or reports that the distance exceeds  $k$ .

First, we describe a simple data structure with  $\tilde{O}(1)$  update time and  $\tilde{O}(k)$  query time. Through considerably more sophisticated techniques, we show that  $\tilde{O}(k)$  update time and  $\tilde{O}(1)$  query time is also achievable. These two solutions likely provide an essentially optimal trade-off for the dynamic  $k$ -mismatch problem with  $m^{\Omega(1)} \leq k \leq \sqrt{m}$ : we prove that, in that case, conditioned on the 3SUM conjecture, one cannot simultaneously achieve  $k^{1-\Omega(1)}$  time for all operations (updates and queries) after  $n^{\Omega(1)}$ -time initialization. For  $k \geq \sqrt{m}$ , the same lower bound excludes achieving  $m^{1/2-\Omega(1)}$  time per operation. This is known to be essentially tight for constant-sized alphabets: already Clifford et al. (STACS 2018) achieved  $\tilde{O}(\sqrt{m})$  time per operation in that case, but their solution for large alphabets costs  $\tilde{O}(m^{3/4})$  time per operation. We improve and extend the latter result by developing a trade-off algorithm that, given a parameter  $1 \leq x \leq k$ , achieves update time  $\tilde{O}(\frac{m}{k} + \sqrt{\frac{mk}{x}})$  and query time  $\tilde{O}(x)$ . In particular, for  $k \geq \sqrt{m}$ , an appropriate choice of  $x$  yields  $\tilde{O}(\sqrt[3]{mk})$  time per operation, which is  $\tilde{O}(m^{2/3})$  when only the trivial threshold  $k = m$  is provided.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Pattern matching

**Keywords and phrases** Pattern matching, Hamming distance, dynamic algorithms

**Digital Object Identifier** 10.4230/LIPIcs.CPM.2022.18

**Funding** *Tomasz Kociumaka*: Partly supported by NSF 1652303, 1909046, and HDR TRIPODS 1934846 grants, and an Alfred P. Sloan Fellowship.

*Przemysław Uznański*: Supported by Polish National Science Centre grant 2019/33/B/ST6/00298.

**Acknowledgements** We are grateful to Ely Porat and Shay Golan for insightful conversations about the dynamic  $k$ -mismatch problem at an early stage of this work.



© Raphaël Clifford, Paweł Gawrychowski, Tomasz Kociumaka, Daniel P. Martin, and Przemysław Uznański;

licensed under Creative Commons License CC-BY 4.0

33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022).

Editors: Hideo Bannai and Jan Holub; Article No. 18; pp. 18:1–18:15



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

The development of dynamic data structures for string problems has become a topic of renewed interest in recent years (see, for example, [2, 3, 4, 5, 10, 11, 13, 15, 20] and references therein). Our focus will be on approximate pattern matching, where the general problem is as follows: Given a pattern of length  $m$  and a longer text of length  $n$ , return the value of a distance function between the pattern and substrings of the text.

We develop new dynamic data structures for a thresholded version of the Hamming distance function, known as the  $k$ -mismatch function. In this setting, we only need to report the Hamming distance if does not exceed  $k$ . The  $k$ -mismatch problem is well studied in the offline setting, where all alignments of the pattern with the text substring that meet this threshold must be found. In 1980s, an  $\mathcal{O}(nk)$ -time algorithm was given [21], and this stood as the record for a over a decade. However, in the last twenty years, significant progress has been made. In a breakthrough result, Amir et al. [6] gave  $\mathcal{O}(n\sqrt{k\log k})$ -time and  $\mathcal{O}(n + \frac{k^3 \log k}{m})$ -time algorithms, which were subsequently improved to  $\mathcal{O}(n \log^{\mathcal{O}(1)} m + \frac{nk^2 \log k}{m})$  time [12],  $\mathcal{O}(n \log^2 m \log \sigma + \frac{nk\sqrt{\log m}}{\sqrt{m}})$  time [16], and finally to  $\mathcal{O}(n + \min(\frac{nk\sqrt{\log m}}{\sqrt{m}}, \frac{nk^2}{m}))$  time [9].

In the dynamic  $k$ -mismatch problem, there are two input strings: a pattern  $P$  of length  $m$  and a text  $T$  of length  $n \geq m$ . For a query at index  $i$ , the data structure must return the Hamming distance between  $P$  and  $T[i \dots i + m]$  if the Hamming distance is less than  $k$ , and  $\infty$  otherwise. The queries can be interspersed with updates of the form  $\text{Update}(S, i, x)$ , which assign  $S[i] := x$ , where  $S$  can be either the pattern or the text. There are two naive approaches for solving the dynamic problem. The first is to rerun a static offline algorithm after each update, and then have constant-time queries. The second is to simply modify the input at each update and compute the Hamming distance naively for each query. Our goal is to perform better than these naive solutions.

We primarily focus on the case when  $n = \Theta(m)$  (in general, one can partition the text into  $\Theta(\frac{n}{m})$  overlapping blocks of length  $\Theta(m)$ ). When  $k = m$  and  $\sigma = n^{\mathcal{O}(1)}$ , known upper bounds and conditional lower bounds match up to a subpolynomial factor: There exists a dynamic data structure with an  $\mathcal{O}(\sqrt{n \log n} \cdot \sigma)$  upper bound for both updates and queries and an almost matching  $n^{1/2 - \Omega(1)}$  lower bound [13] conditioned on the hardness of the online matrix-vector multiplication problem. Although there is no existing work directly on the dynamic  $k$ -mismatch problem we consider, it was shown very recently that a compact representation of all  $k$ -mismatch occurrences can be reported in  $\tilde{\mathcal{O}}(k^2)$  time<sup>1</sup> after each  $\mathcal{O}(\log n)$ -time update [11].

We give three data structures for the dynamic  $k$ -mismatch problem. The first has update time of  $\tilde{\mathcal{O}}(1)$  and a query time of  $\tilde{\mathcal{O}}(k)$ . The main tool we use is the dynamic strings data structure [15] which allows enumerating mismatches in  $\mathcal{O}(\log n)$  time each. The second has update time  $\tilde{\mathcal{O}}(k)$  and a query time of  $\tilde{\mathcal{O}}(1)$ . Here, we build on the newly developed generic solution for the static  $k$ -mismatch problem from [11]. The third data structure, optimized for  $k \geq \sqrt{n}$ , gives a trade-off between update and query times. The overall approach is a lazy rebuilding scheme using the state-of-the-art offline  $k$ -mismatch algorithm. In order to achieve a fast solution, we handle instances with many and few  $2k$ -mismatch occurrences differently. Basing on combinatorial insights developed in the sequence of papers on the offline and streaming versions of the  $k$ -mismatch problem [9, 12, 14, 16, 17], we are able to achieve update time  $\tilde{\mathcal{O}}(\frac{n}{k} + \sqrt{\frac{nk}{x}})$  and query time  $\tilde{\mathcal{O}}(x)$  for any trade-off parameter

<sup>1</sup> The  $\tilde{\mathcal{O}}(\cdot)$  notation suppresses  $\log^{\mathcal{O}(1)} n$  factors.

$x \in [1..k]$  provided at initialization.<sup>2</sup> To put the trade-off complexity in context, we note that, e.g., when  $k = m$ , this allows achieving  $U(n, k) = Q(n, k) = \tilde{O}(n^{2/3})$ , which improves upon an  $\tilde{O}(n^{3/4})$  bound presented in [13] (where only the case of  $k = m$  is considered).

We also show conditional lower bounds which are in most cases within subpolynomial factors of our upper bounds. For the case where the text length is linear in the length of the pattern, we do this by reducing from the 3SUM conjecture [23]. However, in the case that the text is much longer than the pattern, our reduction requires the Online Matrix vector conjecture [18]. Interestingly the lower bound for the superlinear case is asymmetric between the query and update time.

## 2 Preliminaries

In this section, we provide the required basic definitions. We begin with the string distance metric which will be used throughout.

► **Definition 1** (Hamming Distance). *The Hamming distance between two strings  $S, R$  of the same length is defined as  $\text{HD}(S, R) = |\{i : S[i] \neq R[i]\}|$ .*

From this point forward, for simplicity of exposition, we will assume that the pattern is half the length of the text. All our upper bounds are straightforward to generalise to a text whose length is linear in the length of the pattern. In Theorem 27, we show higher lower bounds for the case where the text is much longer than the pattern.

We can now define the central dynamic data structure problem we consider in this paper.

► **Definition 2** (Dynamic  $k$ -Mismatch Problem). *Let  $P$  be a pattern of length  $m$  and  $T$  be a text of length  $n \leq 2m$ . For  $i \in [0..n-m]$ , a query  $\text{Query}(i)$  must return  $\text{HD}(T[i..i+m], P)$  if  $\text{HD}(T[i..i+m], P) \leq k$ , and  $\infty$  otherwise. The queries can be interspersed with updates of the form  $\text{Update}(S, i, x)$  which assign  $S[i] := x$ , where  $S$  can be the pattern or the text.*

For the remainder of the paper, we use  $Q(n, k)$  and  $U(n, k)$  to be the time complexity of Query and Update, respectively. If  $n > 2m$ , then a standard reduction yields  $\mathcal{O}(Q(m, k))$ -time queries,  $\mathcal{O}(U(m, k))$ -time updates in  $T$ , and  $\mathcal{O}(\frac{n}{m}Q(m, k))$ -time updates in  $P$ .

## 3 Upper Bounds

In this section, we provide three solutions of the dynamic  $k$ -mismatch problem. We start with a simple application of dynamic strings resulting in  $\tilde{O}(k)$  query time and  $\tilde{O}(1)$  update time.

The data structure of Gawrychowski et al. [15] maintains a dynamic family  $\mathcal{X}$  of strings of total length  $N$  supporting the following updates:<sup>3</sup>

- Insert to  $\mathcal{X}$  a given string  $S$  (in time  $\mathcal{O}(|S| + \log N)$ ).
- Insert to  $\mathcal{X}$  the concatenation of two strings already in  $\mathcal{X}$  (in time  $\mathcal{O}(\log N)$ ).
- Insert to  $\mathcal{X}$  an arbitrary prefix or suffix of a string already in  $\mathcal{X}$  (in time  $\mathcal{O}(\log N)$ ).

Queries include  $\mathcal{O}(1)$ -time computation of the longest common prefix of two strings in  $\mathcal{X}$ .

<sup>2</sup> Throughout this paper, we denote  $[a..b] = \{i \in \mathbb{Z} : a \leq i \leq b\}$  and  $[a..b) = \{i \in \mathbb{Z} : a \leq i < b\}$ .

<sup>3</sup> This data structure is Las-Vegas randomized, and the running times are valid with high probability with respect to  $N$ . A deterministic version, using [1] and deterministic dynamic dictionaries, has an  $\mathcal{O}(\log N)$ -factor overhead in the running times, which translates to an  $\mathcal{O}(\log n)$ -factor overhead in the query and update times of all our randomized algorithms for the dynamic  $k$ -mismatch problem.

► **Theorem 3.** *There exists a Las-Vegas randomized algorithm for the dynamic  $k$ -mismatch problem satisfying  $U(n, k) = \mathcal{O}(\log n)$  and  $Q(n, k) = \mathcal{O}(k \log n)$  with high probability.*

**Proof.** We maintain a dynamic string collection  $\mathcal{X}$  of [15] containing  $P$  and  $T$ . Given that a string  $S'$  resulting from setting  $S[i] := x$  in a string  $S \in \mathcal{X}$  is the concatenation of a prefix  $S[0..i)$ , the new character  $x$ , and a suffix  $S[i+1..|S|)$ , it is straightforward to construct  $S'$  with  $\mathcal{O}(1)$  auxiliary strings added to  $\mathcal{X}$ . Hence, we implement an update in  $\mathcal{O}(\log n)$  time.

Armed with this tool, we perform dynamic  $k$ -mismatch queries by so-called “kangaroo jumps” [21]. That is, we align the pattern with  $T[i..i+m)$ , where  $i$  is the query position in the text  $T$ , and we repeatedly extend the match we have found so far until we reach a fresh mismatch. Each longest common extension query can be implemented in  $\mathcal{O}(\log n)$  time. For this, we extract the relevant suffixes of  $P$  and  $T$  (we insert them to  $\mathcal{X}$  in  $\mathcal{O}(\log n)$  time each) and ask for their longest common prefix (which costs  $\mathcal{O}(1)$  time). As we stop once  $k+1$  mismatches have been found or once we have reached the end of the text or pattern, the total query time is  $\mathcal{O}(k \log n)$ . ◀

### 3.1 Faster Queries, Slower Updates

Given the result above, a natural question is whether there exists an approach with an efficient query algorithm, in return for a slower update algorithm. We answer affirmatively in this section based on a recent work of Charalampopoulos et al. [11].

#### 3.1.1 The PILLAR model

Charalampopoulos et al. [11] developed a generic static algorithm for the  $k$ -mismatch problem. They formalized their solution using an abstract interface, called the *PILLAR model*, which captures certain primitive operations that can be implemented efficiently in all settings considered in [11]. Thus, we bound the running times in terms of PILLAR operations – if the algorithm uses more time than PILLAR operations, we also specify the extra running time.

In the PILLAR model, we are given a family of strings  $\mathcal{X}$  for preprocessing. The elementary objects are fragments  $X[\ell..r)$  of strings  $X \in \mathcal{X}$ . Initially, the model provides access to each  $X \in \mathcal{X}$  interpreted as  $X[0..|X|)$ . Other fragments can be obtained through an **Extract** operation.

- **Extract**( $S, \ell, r$ ): Given a fragment  $S$  and positions  $0 \leq \ell \leq r \leq |S|$ , extract the (sub)fragment  $S[\ell..r)$ , which is defined as  $X[\ell'+\ell.. \ell'+r)$  if  $S = X[\ell'..r')$  for  $X \in \mathcal{X}$ . Furthermore, the following primitive operations are supported in the PILLAR model:
- **LCP**( $S, T$ ): Compute the length of the longest common prefix of  $S$  and  $T$ .
- **LCP<sup>R</sup>**( $S, T$ ): Compute the length of the longest common suffix of  $S$  and  $T$ .
- **IPM**( $P, T$ ): Assuming that  $|T| \leq 2|P|$ , compute the occurrences of  $P$  in  $T$ , i.e.,  $\text{Occ}(P, T) = \{i \in [0..|T| - |P|] : P = T[i..i + |P|)\}$  represented as an arithmetic progression.
- **Access**( $S, i$ ): Retrieve the character  $S[i]$ .
- **Length**( $S$ ): Compute the length  $|S|$  of the string  $S$ .

Among several instantiations of the model, Charalampopoulos et al. [11, Section 7.3] showed that the primitive PILLAR operations can be implemented in  $\mathcal{O}(\log^2 N)$  time on top of the data structure for dynamic strings [15], which we recalled above. Consequently, we are able to maintain two dynamic strings  $P$  and  $T$  subject to character substitutions, achieving  $\mathcal{O}(\log^2 n)$ -time elementary PILLAR operations and  $\mathcal{O}(\log n)$ -time updates.

► **Corollary 4.** *Let  $T$  be a dynamic string of length  $n$  and  $P$  be a dynamic string of length  $m \leq n$ , both of which can be updated via substitutions of single characters. There exists a Las-Vegas randomized data structure supporting the PILLAR operations on  $\mathcal{X} = \{T, P\}$  in  $\mathcal{O}(\log^2 n)$  time w.h.p. and updates in  $\mathcal{O}(\log n)$  time w.h.p.*

### 3.1.2 The Static $k$ -Mismatch Problem

The (static)  $k$ -mismatch problem consists in computing  $\text{Occ}_k(P, T) = \{i \in [0..n - m] : \text{HD}(P, T[i..i + m]) \leq k\}$ , with each position  $i \in \text{Occ}_k(P, T)$  reported along with the corresponding Hamming distance  $d_i := \text{HD}(P, T[i..i + m])$ . Charalampopoulos et al. [11, Theorem 3.1 and Corollary 3.5] proved that  $\text{Occ}_k(P, T)$  admits a compact representation: this set can be decomposed into  $\mathcal{O}(\frac{n}{m} \cdot k^2)$  disjoint arithmetic progressions so that occurrences in a single progression share the same Hamming distance  $d_i$ . Moreover, all the non-trivial progressions (i.e., progressions with two or more terms) share the same difference. The following algorithm gives this compact representation on the output.

► **Theorem 5** ([11, Main Theorem 8]). *There exists a PILLAR-model algorithm that, given a pattern  $P$  of length  $m$ , a text  $T$  of length  $n \geq m$ , and a positive integer  $k \leq m$ , solves the  $k$ -mismatches problem in  $\mathcal{O}(\frac{n}{m} \cdot k^2 \log \log k)$  time using  $\mathcal{O}(\frac{n}{m} \cdot k^2)$  PILLAR operations.*

### 3.1.3 Warm-Up Algorithm

Intuitively, the algorithm of Theorem 5 precomputes the answers to all queries  $\text{Query}(i)$  with  $i \in [0..n - m]$ . Hence, a straightforward solution to the dynamic  $k$ -mismatch problem would be to maintain the data structure of Corollary 4, use the algorithm of Theorem 5 after each update, and then retrieve the precomputed answers for each query asked. The data structure described below follows this strategy, making sure that the compact representation of  $\text{Occ}_k(P, T)$  is augmented with infrastructure for efficient random access.

► **Proposition 6.** *There exists a Las-Vegas algorithm for the dynamic  $k$ -mismatch problem satisfying  $U(n, k) = \mathcal{O}(k^2 \log^2 n)$  and  $Q(n, k) = \mathcal{O}(\log \log n)$  with high probability.*

**Proof.** We maintain a PILLAR-model implementation of  $\mathcal{X} = \{P, T\}$  using Corollary 4; this costs  $\mathcal{O}(\log n)$  time per update and provides  $\mathcal{O}(\log^2 n)$ -time primitive PILLAR operations.

Following each update, we use Theorem 5 so that a space-efficient representation of  $\text{Occ}_k(P, T)$  is computed in  $\mathcal{O}(k^2 \log^2 n)$  time (recall that  $m = \Theta(n)$ ). This output is then post-processed as described below. Let  $q$  be the common difference of non-trivial arithmetic progression forming  $\text{Occ}_k(P, T)$ ; we set  $q = 1$  if all progressions are trivial. Consider the indices  $i \in [0..n - m]$  ordered by  $(i \bmod q, i)$ , that is, first by the remainder modulo  $q$  and then by the index itself. In this ordering, each arithmetic progression contained in the output  $\text{Occ}_k(P, T)$  yields a contiguous block of indices  $i$  with a common finite answer to queries  $\text{Query}(i)$ . The goal of post-processing is to store the sequence of answers using run-length encoding (with run boundaries kept in a predecessor data structure). This way, for each of the  $\mathcal{O}(k^2)$  arithmetic progressions in  $\text{Occ}_k(P, T)$ , the corresponding answers  $\text{Query}(i)$  can be set in  $\mathcal{O}(\log \log n)$  time to the common value  $d_i$  reported along with the progression. In total, the post-processing time is therefore  $\mathcal{O}(k^2 \log \log n)$ .

At query time, any requested value  $\text{Query}(i)$  can be retrieved in  $\mathcal{O}(\log \log n)$  time. ◀

### 3.1.4 Structural Insight

In order to improve the update time, we bring some of the combinatorial insight from [11].

A string is *primitive* if it is not a string power with an integer exponent strictly greater than 1. For a non-empty string  $Q$ , we denote by  $Q^\infty$  an infinite string obtained by concatenating infinitely many copies of  $Q$ . For an arbitrary string  $S$ , we further set  $\text{HD}(S, Q^*) = \text{HD}(S, Q^\infty[0..|S|])$ . In other words, the  $\text{HD}(\cdot, \cdot^*)$  function generalizes  $\text{HD}(\cdot, \cdot)$  in that the second string is cyclically extended to match the length of the first one. We use the same convention to define  $M(S, Q^*) = \{i : S[i] \neq Q^\infty[i]\} = \{i : S[i] \neq Q[i \bmod |Q|]\}$ .

► **Proposition 7** ([11, Theorems 3.1 and 3.2]). *Let  $P$  be a pattern of length  $m$ , let  $T$  be a text of length  $n \leq \frac{3}{2}m$ , and let  $k \leq m$  be a positive integer. At least one of the following holds:*

1. *The number of  $k$ -mismatch occurrences of  $P$  in  $T$  is  $|\text{Occ}_k(P, T)| \leq 864k$ .*
  2. *There is a primitive string  $Q$  of length  $|Q| \leq \frac{m}{128k}$  such that  $\text{HD}(P, Q^*) < 2k$ .*
- Moreover, if  $\text{Occ}_k(P, T) \neq \emptyset$  and (2) holds, then a fragment  $T' = T[\min \text{Occ}_k(P, T) .. m + \max \text{Occ}_k(P, T)]$  satisfies  $\text{HD}(T', Q^*) < 6k$  and every position in  $\text{Occ}_k(P, T')$  is a multiple of  $|Q|$ .*

We also need a characterization of the values  $\text{HD}(P, T'[j|Q| .. m + j|Q|])$ .

► **Proposition 8** ([11, Lemma 3.3 and Claim 3.4]). *Let  $P$  be a pattern of length  $m$ , let  $T$  be a text of length  $n$ , and let  $k \leq m$  be a positive integer. For any non-empty string  $Q$  and non-negative integer  $j \leq \frac{n-m}{|Q|}$ , we have*

$$\text{HD}(P, T[j|Q| .. m + j|Q|]) = |M(P, Q^*)| + |M(T, Q^*) \cap [j|Q| .. m + j|Q|]| - \mu_j,$$

where

$$\mu_j = \sum_{\rho \in M(P, Q^*), \tau \in M(T, Q^*) : \tau = j|Q| + \rho} 2 - \text{HD}(T[\tau], P[\rho]).$$

### 3.1.5 Improved Solution

The idea behind achieving  $\mathcal{O}(k \log^2 n)$  update time is to run Theorem 5 once every  $k$  updates, but with a doubled threshold  $2k$  instead of  $k$ . The motivation behind this choice of parameters is that if the current instance  $P, T$  is obtained by up to  $k$  substitutions from a past instance  $\bar{P}, \bar{T}$ , then  $\text{HD}(P, \bar{P}) + \text{HD}(T, \bar{T}) \leq k$  yields  $\text{Occ}_k(P, T) \subseteq \text{Occ}_{2k}(\bar{P}, \bar{T})$ . Consequently, the algorithm may safely return  $\infty$  while answering  $\text{Query}(i)$  for any position  $i \notin \text{Occ}_{2k}(\bar{P}, \bar{T})$ .

If the application of Theorem 5 identifies few  $2k$ -mismatch occurrences, then we maintain the Hamming distances  $d_i$  at these positions throughout the  $k$  subsequent updates. Otherwise, we identify  $Q$  and  $T'$ , as defined in Proposition 7, as well as the sets  $M(P, Q^*)$ ,  $M(T', Q^*)$ , and the values  $\mu_j$  of Proposition 8 so that the distances  $\text{HD}(P, T'[j|Q| .. m + j|Q|])$  can be retrieved efficiently.

The latter task requires extending Theorem 5 so that the string  $Q$  and the sets  $M(P, Q^*)$ ,  $M(T', Q^*)$  can be constructed whenever there are many  $k$ -mismatch occurrences.

► **Lemma 9.** *There exists a PILLAR-model algorithm that, given a pattern  $P$  of length  $m$ , a text  $T$  of length  $n \leq \frac{3}{2}m$ , and a positive integer  $k \leq m$ , returns  $\text{Occ}_k(P, T)$  along with the corresponding Hamming distances provided that  $|\text{Occ}_k(P, T)| \leq 864k$ , or, otherwise, returns the fragment  $T' = T[\min \text{Occ}_k(P, T) .. m + \max \text{Occ}_k(P, T)]$ , a string  $Q$  such that  $\text{HD}(P, Q^*) < 2k$ ,  $\text{HD}(T', Q^*) < 6k$ , and  $\text{Occ}_k(P, T')$  consists of multiples of  $|Q|$ , and sets  $M(P, Q^*)$ ,  $M(T', Q^*)$ . The algorithm takes  $\mathcal{O}(k^2 \log \log k)$  time plus  $\mathcal{O}(k^2)$  PILLAR operations.*

**Proof.** First, we use Theorem 5 in order to construct  $\text{Occ}_k(P, T)$  in a compact representation as  $\mathcal{O}(k^2)$  arithmetic progressions. Based on this representation, both  $|\text{Occ}_k(P, T)|$  and  $T'$  can be computed in  $\mathcal{O}(k^2)$  time. If  $|\text{Occ}_k(P, T)| \leq 864k$ , then  $\text{Occ}_k(P, T)$  is converted to a plain representation (with each position reported explicitly along with the corresponding Hamming distance). Otherwise, we use the  $\text{Analyze}(P, k)$  procedure of [11, Lemma 4.4]. This procedure costs  $\mathcal{O}(k)$  time in the PILLAR model, and it detects a structure within the pattern  $P$  that can be of one of three types. A possible outcome includes a primitive string  $Q$  such that  $|Q| \leq \frac{m}{128k}$  and  $\text{HD}(P, Q^*) < 8k$ . Moreover, the existence of a structure of either of the other two types contradicts  $|\text{Occ}_k(P, T)| \leq 864k$  (due to [11, Lemmas 3.8 and 3.11]), and so does  $2k \leq \text{HD}(P, Q^*) < 8k$  (due to [11, Lemma 3.14]). Consequently, we are guaranteed to obtain a primitive string  $Q$  such that  $|Q| \leq \frac{m}{128k}$  and  $\text{HD}(P, Q^*) < 2k$ , which are precisely the conditions in the second case of Proposition 7. Thus, we conclude that  $\text{HD}(T', Q^*) < 6k$  and that  $\text{Occ}_k(P, T')$  consists of multiples of  $|Q|$ . It remains to report  $M(P, Q^*)$  and  $M(T', Q^*)$ . For this task, we employ [11, Corollary 4.2], whose time cost in the PILLAR model is proportional to the output size, i.e.,  $\mathcal{O}(k)$  for both instances. ◀

We are now ready to describe the dynamic algorithm based on the intuition above. Initially, we only improve the *amortized* query time from  $\mathcal{O}(k^2 \log^2 n)$  to  $\mathcal{O}(k \log^2 n)$ .

► **Proposition 10.** *There exists a Las-Vegas randomized algorithm for the dynamic  $k$ -mismatch problem satisfying  $Q(n, k) = \mathcal{O}(\log \log n)$  and  $U(n, k) = \mathcal{O}(k + \log n)$  with high probability, except that every  $k$ th update costs  $\mathcal{O}(k^2 \log^2 n)$  time w.h.p.*

**Proof.** The algorithm logically partitions its runtime into epochs, with  $k$  updates in each epoch. The first update in every epoch costs  $\mathcal{O}(k^2 \log^2 n)$  time, and the remaining updates cost  $\mathcal{O}(k + \log n)$  time. A representation of  $\mathcal{X} = \{P, T\}$  supporting the PILLAR operations (Corollary 4) is maintained throughout the execution of the algorithm, while the remaining data is destroyed after each epoch.

Once the arrival of an update marks the beginning of a new epoch, we run the algorithm of Lemma 9 with a doubled threshold  $2k$ . This procedure costs  $\mathcal{O}(k^2 \log^2 n)$  time, and it may have one of two types of outcome.

The first possibility is that it returns a set  $O := \text{Occ}_{2k}(P, T)$  of up to  $1728k$  positions, with the Hamming distance  $d_i := \text{HD}(P, T[i..i+m])$  reported along with each position  $i \in O$ . Since  $d_i > 2k$  for  $i \notin O$  and any update may decrease  $d_i$  by at most one, we are guaranteed that  $\text{Query}(i) = \infty$  can be returned for  $i \notin O$  for the duration of the epoch. Consequently, the algorithm only maintains  $d_i$  for  $i \in O$ . For each of the subsequent updates, the algorithm iterates over  $i \in O$  and checks if  $d_i$  needs to be changed: If the update involves  $P[j]$ , then both the old and the new value of  $P[j]$  are compared against  $T[i+j]$ . Similarly, if the update involves  $T[j]$  and  $j \in [i..i+m)$ , then both the old and the new value of  $T[j]$  are compared against  $P[i-j]$ . Thus, the update time is  $\mathcal{O}(k)$  and the query time is  $\mathcal{O}(1)$ .

The second possibility is that the algorithm of Lemma 9 results in a fragment  $T' = T[\ell..r)$ , a string  $Q$ , and the mismatching positions  $M(P, Q^*)$  and  $M(T', Q^*)$ . We are then guaranteed that each  $2k$ -mismatch occurrence of  $P$  in  $T$  starts at a position  $i \in [\ell..r-m]$  congruent to  $\ell$  modulo  $|Q|$ . We call these positions *relevant*. As in the previous case,  $\text{Query}(i) = \infty$  can be returned for irrelevant  $i$  for the duration of the epoch. The Hamming distances  $d_i$  at relevant positions are computed using Proposition 8. For this, we maintain  $M(P, Q^*)$ ,  $M(T', Q^*)$ , and all non-zero values  $\mu_j$  for  $j \in [0.. \lfloor \frac{r-\ell-m}{|Q|} \rfloor]$ . Moreover,  $M(T', Q^*)$  is stored in a predecessor data structure, and each element of  $M(T', Q^*)$  maintains its rank in this set. Every subsequent update affects at most one element of  $M(P, Q^*)$  or  $M(T', Q^*)$ , so these sets can be updated in  $\mathcal{O}(1)$  time. Maintaining the predecessor data structure costs further

$\mathcal{O}(\log \log n)$  time, and maintaining the ranks costs up to  $\mathcal{O}(\text{HD}(T', Q^*))$  time. In order to update the values  $\mu_j$ , we proceed as follows. If the update involves a character  $P[\rho]$ , we iterate over  $\tau \in M(T', Q^*)$ . If  $j = \tau - \rho|Q|$  is an integer between 0 and  $\frac{r-\ell-m}{|Q|}$ , we may need to update the entry  $\mu_j$  (which costs constant time). An update involving  $T[\ell + \tau]$  is processed in a similar way. Overall, the update time is  $\mathcal{O}(\log n + \text{HD}(T', Q^*) + \text{HD}(P, Q^*)) = \mathcal{O}(\log n + k)$  because  $\text{HD}(P, Q^*) + \text{HD}(T', Q^*) < 2k + 6k + k = 9k$  holds for the duration of the epoch.

As for the query  $\text{Query}(i)$ , we return  $\infty$  if  $i$  is irrelevant, i.e.,  $i < \ell$ ,  $i > r - m$ , or  $i \not\equiv \ell \pmod{|Q|}$ . Otherwise, we set  $j = \frac{i-\ell}{|Q|}$  and, according to Proposition 8, return  $|M(P, Q^*)| + |M(T', Q^*) \cap [j|Q|..j|Q| + m]| - \mu_j$ . The second term is determined in  $\mathcal{O}(\log \log n)$  time using the predecessor data structure on top of  $M(T', Q^*)$  as well as the rank stored for each element of this set.  $\blacktriangleleft$

Finally, we show how to achieve *worst-case*  $\mathcal{O}(k \log^2 n)$  update time.

► **Theorem 11.** *There exists a Las-Vegas randomized algorithm for the dynamic  $k$ -mismatch problem satisfying  $Q(n, k) = \mathcal{O}(\log \log n)$  and  $U(n, k) = \mathcal{O}(k \log^2 n)$  with high probability.*

**Proof.** We maintain two instances of the algorithm of Proposition 10, with updates forwarded to both instances, but queries forwarded to a single instance that is currently *active*.

The algorithm logically partitions its runtime into epochs, with  $\frac{1}{2}k$  updates in each epoch. For the two instances, the time-consuming updates are chosen to be the first updates of every even and odd epoch, respectively. Once an instance has to perform a time-consuming update, it becomes inactive (it buffers the subsequent updates and cannot be used for answering queries) and stays inactive for the duration of the epoch. The work needed to perform the time-consuming update is spread across the time allowance for the first half of the epoch, with the time allowance for the second half of the epoch used in order to clear the accumulated backlog of updates (by processing updates at a doubled rate). During this epoch, the other (active) instance processes updates and queries as they arrive in  $\mathcal{O}(k \log^2 n)$  and  $\mathcal{O}(\log \log n)$  worst-case time, respectively.  $\blacktriangleleft$

### 3.2 Trade-off between Update Time and Query Time

The next natural question is the existence of a trade-off between the run-times of Theorems 3 and 11. Due to Theorem 23 (in Section 4), the answer is likely negative for  $k \ll \sqrt{n}$ . Nevertheless, for  $k \gg \sqrt{n}$ , the trade-off presented below simultaneously achieves  $Q(n, k), U(n, k) = k^{1-\Omega(1)}$ .

We first recall some combinatorial properties originating from previous work on the  $k$ -mismatch problem [9, 12, 14, 17]. The description below mostly follows [17, Section 3].

► **Definition 12** ([12]). *Let  $X$  be a string and let  $d$  be a non-negative integer. A positive integer  $\rho \leq |X|$  is a  $d$ -period of  $X$  if  $\text{HD}(X[\rho..|X|], X[0..|X| - \rho]) \leq d$ .*

Recall that  $\text{Occ}_k(P, T) = \{i : \text{HD}(P, T[i..i + m]) \leq k\}$  for a pattern  $P$  and text  $T$ .

► **Lemma 13** ([12]). *If  $i, i' \in \text{Occ}_k(P, T)$  are distinct, then  $\rho := |i' - i|$  is a  $2k$ -period of  $P$ . Moreover, if  $n \leq 2m$ , then  $\rho$  is a  $(8k + \rho)$ -period of  $T[\min \text{Occ}_k(P)..m + \max \text{Occ}_k(P)]$ .*

Recall that the  $L_0$ -norm of a function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  defined as  $\|f\|_0 = |\{x : f(x) \neq 0\}|$ . The *convolution* of two functions  $f, g : \mathbb{Z} \rightarrow \mathbb{Z}$  with finite  $L_0$ -norms is a function  $f * g : \mathbb{Z} \rightarrow \mathbb{Z}$  such that

$$[f * g](i) = \sum_{j \in \mathbb{Z}} f(j) \cdot g(i - j).$$



For a string  $X$  over  $\Sigma$  and a symbol  $c \in \Sigma$ , the *characteristic function* of  $X$  and  $c$  is  $X_c : \mathbb{Z} \rightarrow \{0, 1\}$  such that  $X_c(i) = 1$  if and only if  $X[i] = c$ . For a string  $X$ , let  $X^R$  denote  $X$  reversed. The *cross-correlation* of strings  $X$  and  $Y$  over  $\Sigma$  is a function  $X \otimes Y : \mathbb{Z} \rightarrow \mathbb{Z}$  such that

$$X \otimes Y = \sum_{c \in \Sigma} X_c * Y_c^R.$$

► **Fact 14** ([14, Fact 7.1]). *For  $i \in [m-1..n]$ , we have  $[T \otimes P](i) = |P| - \text{HD}(P, T(i-m..i))$ . For  $i < 0$  and for  $i \geq m+n$ , we have  $[T \otimes P](i) = 0$ .*

By Fact 14,  $[T \otimes P](i+m-1)$  suffices to compute  $\text{HD}(P, T[i..i+m])$  for  $i \in [0..n-m]$ . The *backward difference* of a function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  due to  $\rho \in \mathbb{Z}_+$  is  $\Delta_\rho[f](i) = f(i) - f(i-\rho)$ .

► **Observation 15** ([14, Observation 7.2]). *If a string  $X$  has a  $d$ -period  $\rho$ , then*

$$\sum_{c \in \Sigma} \|\Delta_\rho[X_c]\|_0 \leq 2(d + \rho).$$

Our computation of  $T \otimes P$  is based on the following lemma:

► **Lemma 16** (See [17, Lemma 6]). *For every pattern  $P$ , text  $T$ , and positive integer  $\rho$ , we have  $\Delta_\rho[\Delta_\rho[T \otimes P]] = \sum_{c \in \Sigma} \Delta_\rho[T_c] * \Delta_\rho[P_c^R]$ . Consequently, for every  $i \in \mathbb{Z}$ ,*

$$[T \otimes P](i) = \sum_{j=0}^{\infty} (j+1) \cdot \left[ \sum_{c \in \Sigma} \Delta_\rho[T_c] * \Delta_\rho[P_c^R] \right] (i - j\rho).$$

► **Theorem 17.** *There exists a deterministic algorithm for the dynamic  $k$ -mismatch problem with  $U(n, k) = \mathcal{O}\left(\sqrt{\frac{nk}{x}} + \frac{n}{k}\right)$  and  $Q(n, k) = \tilde{\mathcal{O}}(x)$ , where  $x$  is a trade-off parameter that can be set in  $[1..k]$ .*

**Proof.** We solve the problem using a lazy rebuilding scheme similar to that in the proof of Theorem 11. Hence, we can afford update time  $\tilde{\mathcal{O}}(n + k\sqrt{n})$  every  $k$  updates. Thus, if an incoming update marks the beginning of a new epoch (lasting for  $k$  updates), we run a (static)  $2k$ -mismatch algorithm [9, 16], resulting in  $O := \text{Occ}_{2k}(P, T)$  and the Hamming distances  $d_i = \text{HD}(P, T[i..i+m])$  for each  $i \in O$ . This takes  $\tilde{\mathcal{O}}(n + k\sqrt{n})$  time. As in the proof of Proposition 10, since  $\text{Occ}_k(P, Q) \subseteq O$  holds for the duration of the epoch, we can safely return  $\infty$  for  $\text{Query}(i)$  with  $i \notin O$ . We distinguish two cases.

$|O| \leq \frac{n}{k}$ : We maintain the distances  $d_i$  for  $i \in O$ . As noted above,  $\text{Occ}_k(P, T) \subseteq O$  even after  $k$  updates. We now observe that any update requires only updating the mismatches for every element of  $O$ , with  $\mathcal{O}(1)$  cost per element and  $\mathcal{O}(\frac{n}{k})$  total; the queries are handled by finding the answer stored for  $i \in O$ , at  $\tilde{\mathcal{O}}(1)$  cost.

$|O| > \frac{n}{k}$ : We set  $\rho$  to be the distance between two closest elements of  $O$ ; we have  $\rho \leq k$  due to  $|O| > \frac{n}{k}$ . By Lemma 13,  $\rho$  is a  $4k$ -period of  $P$  and a  $17k$ -period of  $T' := T[\min O..m + \max O]$ . Moreover,  $\text{Occ}_k(P, T) \subseteq O \subseteq [\min O..m + \max O]$  holds for the duration of the epoch, so all  $k$ -mismatch occurrences of  $P$  in  $T$  remain contained in  $T'$ .

We have thus reduced our problem to answering queries and performing updates for  $P$  and  $T'$ . Moreover, we have a positive integer  $\rho \leq k$  which is initially a  $4k$ -period of  $P$  and a  $17k$ -period of  $T'$ , and, after  $k$  updates, it remains a  $6k$ -period of  $P$  and  $19k$ -period of  $T'$ . Let us define the *weight* of  $c \in \Sigma$  as  $\|\Delta_\rho[P_c^R]\|_0 + \|\Delta_\rho[T'_c]\|_0$ ; by Observation 15, the total weight across  $c \in \Sigma$  remains  $\mathcal{O}(k)$ .

## 18:10 The Dynamic $k$ -Mismatch Problem

We proceed as follows. We maintain  $\Delta_\rho(P_c^R) * \Delta_\rho(T'_c)$  for each letter  $c \in \Sigma$  separately, and the sum  $\Delta_\rho[\Delta_\rho[T \otimes P]] = \sum_{c \in \Sigma} \Delta_\rho(P_c^R) * \Delta_\rho(T'_c)$ . For each remainder  $i \bmod \rho$ , the values  $\Delta_\rho[\Delta_\rho[T \otimes P]](i)$  are stored in a data structure that allows queries for prefix sums (both unweighted and weighted by  $\lfloor i/\rho \rfloor$ ) so that  $[T \otimes P](i)$  can be retrieved efficiently using Lemma 16. Every update to  $P$  or  $T$  incurs updates to  $\Delta_\rho(P_c^R)$  or  $\Delta_\rho(T'_c)$ , in  $\mathcal{O}(1)$  places in total (two for each letter involved in the substitution). We buffer the updates to those convolutions of (potentially) sparse functions during subepochs of  $x$  updates, and then we recompute values of  $\Delta_\rho(P_c^R) * \Delta_\rho(T'_c)$  amortized during the next  $x$  updates. We fix a threshold value  $t$  (specified later), and iterate through letters  $c \in \Sigma$ .

- If a letter  $c$  had weight at least  $t$  or accumulated at least  $t$  updates, we recompute the corresponding convolution from scratch, at the cost of  $\tilde{\mathcal{O}}(n)$  time per each such *heavy* letter.
- Otherwise, updates are processed one by one, at the cost of  $\tilde{\mathcal{O}}(t)$  time per update.

There are  $\mathcal{O}(\frac{k+x}{t})$  heavy letters, which is  $\tilde{\mathcal{O}}(\frac{k}{t})$  since  $x \leq k$ . Thus, the total cost  $\tilde{\mathcal{O}}(\frac{nk}{t} + xt)$  is minimized when  $t = \sqrt{\frac{nk}{x}}$  and gives  $\tilde{\mathcal{O}}(\sqrt{nkx})$  time per subepoch, or  $\tilde{\mathcal{O}}(\sqrt{\frac{nk}{x}})$  time per update.

To perform queries, we retrieve  $[T \otimes P](i + m - 1)$  using Lemma 16 and the data structure maintaining  $\Delta_\rho[\Delta_\rho[T \otimes P]]$  to recover the number of matches last time we stored the convolutions. Next, we scan through the list of at most  $2x$  updates to potentially update the answer. ◀

To put the trade-off complexity in context, we note that e.g., when  $k = m$ , it is possible to achieve  $U(n, k), Q(n, k) = \tilde{\mathcal{O}}(n^{2/3})$ . This improves over  $\tilde{\mathcal{O}}(n^{3/4})$  presented in [13].

### 4 Lower Bounds

In this section, we give conditional lower bounds for the dynamic  $k$ -mismatch problem based on the 3SUM conjecture [23]. For the 3SUM problem, we use the following definition.

► **Definition 18** (3SUM Problem). *Given three sets  $A, B, C \subseteq [-N..N]$  of total size  $|A| + |B| + |C| = n$ , decide whether there exist  $a \in A, b \in B, c \in C$  such that  $a + b + c = 0$ .*

Henceforth, we consider algorithms for the word RAM model with  $w$ -bit machine words, where  $w = \Omega(\log N)$ . In this model, there is a simple  $\mathcal{O}(n^2)$ -time solution for the 3SUM problem. This can be improved by log factors [7], with the current record being  $\mathcal{O}((n^2/\log^2 n)(\log \log n)^{\mathcal{O}(1)})$  time [8].

► **Conjecture 19** (3SUM Conjecture). *For every constant  $\varepsilon > 0$ , there is no Las-Vegas randomized algorithm solving the 3SUM problem in  $\mathcal{O}(n^{2-\varepsilon})$  expected time.*

As a first step, we note that the 3SUM problem remains hard even if we allow for polynomial-time preprocessing of  $A$ . The following reduction is based on [22, Theorem 13].

► **Lemma 20.** *Suppose that, for some constants  $d \geq 2$  and  $\varepsilon > 0$ , there exists an algorithm that, after preprocessing integers  $n, N \in \mathbb{Z}_+$  and a set  $A \subseteq [-N..N]$  in  $\mathcal{O}(n^d)$  expected time, given sets  $B, C \subseteq [-N..N]$  of total size  $|A| + |B| + |C| \leq n$ , solves the underlying instance of the 3SUM problem in expected  $\mathcal{O}(n^{2-\varepsilon})$  time. Then, the 3SUM conjecture fails.*

**Proof.** We shall demonstrate an algorithm solving the 3SUM problem in  $\mathcal{O}(n^{2-\hat{\varepsilon}})$  time, where  $\hat{\varepsilon} = \min(\frac{1}{2}, \frac{\varepsilon}{2(d-1)}) > 0$ . Let  $g = \lceil n^{\frac{d-1.5}{d-1}} \rceil$ . We construct a decomposition  $A = \bigcup_{i=1}^g A_i$  into disjoint subsets such that  $|A_i| \leq \lceil \frac{1}{g}|A| \rceil$  and  $\max A_i < \min A_{i'}$  hold for  $i, i' \in [1..g]$  with  $i < i'$ . Similarly, we also decompose  $B = \bigcup_{j=1}^g B_j$  and  $C = \bigcup_{k=1}^g C_k$ .

Next, we construct  $T = \{(i, j, k) \in [1..g]^3 : \min A_i + \min B_j + \min C_k \leq 0 \leq \max A_i + \max B_j + \max C_k\}$ . Observe that if  $a + b + c = 0$  for  $(a, b, c) \in A \times B \times C$ , then the triple  $(i, j, k) \in [1..g]^3$  satisfying  $(a, b, c) \in A_i \times B_j \times C_k$  clearly belongs to  $T$ . Moreover,  $T$  can be constructed in  $\mathcal{O}(g^2 \log g + |T|)$  time by performing a binary search over  $k \in [1..g]$  for all  $(i, j) \in [1..g]^2$ . To provide a worst-case bound on this running time, we shall prove that  $|T| = \mathcal{O}(g^2)$ . For this, let us define the *domination* order  $\prec$  on  $[1..g]^3$  so that  $(i, j, k) \prec (i', j', k')$  if and only if  $i < i'$ ,  $j < j'$ , and  $k < k'$ . Observe that  $T$  is an  $\prec$ -antichain and that  $[1..g]^3$  can be covered with  $\mathcal{O}(g^2)$   $\prec$ -chains. Hence,  $|T| = \mathcal{O}(g^2)$  holds as claimed.

Let  $\hat{n} := \left\lceil \frac{1}{g}|A| \right\rceil + \left\lceil \frac{1}{g}|B| \right\rceil + \left\lceil \frac{1}{g}|C| \right\rceil = \mathcal{O}\left(\frac{n}{g}\right)$ . We preprocess  $(\hat{n}, N, A_i)$  for each  $i \in [1..g]$ , at the cost of  $\mathcal{O}(g\hat{n}^d) = \mathcal{O}\left(\frac{n^d}{g^{d-1}}\right) = \mathcal{O}\left(\frac{n^d}{n^{d-1.5}}\right) = \mathcal{O}(n^{1.5})$  time. Then, for each triple  $(i, j, k) \in T$ , we solve the underlying instance of the 3SUM problem, at the cost of  $\mathcal{O}(g^2 \hat{n}^{2-\varepsilon}) = \mathcal{O}\left(g^2 \frac{n^{2-\varepsilon}}{g^{2-\varepsilon}}\right) = \mathcal{O}(n^{2-\varepsilon} g^\varepsilon) = \mathcal{O}\left(n^{2-\varepsilon+\varepsilon \frac{d-1.5}{d-1}}\right) = \mathcal{O}\left(n^{2-\frac{\varepsilon}{2(d-1)}}\right)$  expected time in total. As noted above, it suffices to return YES if and only if at least one of these calls returns YES.  $\blacktriangleleft$

Our lower bounds also rely on the following variant of the 3SUM problem.

► **Definition 21 (3SUM<sup>+</sup> Problem).** *Given three sets  $A, B, C \subseteq [-N..N]$  of total size  $|A| + |B| + |C| = n$ , report all  $c \in C$  such that  $a + b + c = 0$  for some  $a \in A$  and  $b \in B$ .*

The benefit of using 3SUM<sup>+</sup> is that it remains hard for  $N \geq n^{2+\Omega(1)}$  (as shown in [19]); in comparison, regular 3SUM is known to be hard only for  $N \geq n^3$ . The following proposition generalizes the results of [19] (allowing for preprocessing of  $A$ ); its proof relies on the techniques of [7].

► **Proposition 22.** *Suppose that, for some constants  $d \geq 2$  and  $\varepsilon, \delta > 0$ , there exists an algorithm that, after  $\mathcal{O}(n^d)$ -time preprocessing of integers  $n, N \in \mathbb{Z}_+$ , with  $N \leq n^{2+\delta}$ , and a set  $A \subseteq [-N..N]$ , given sets  $B, C \subseteq [-N..N]$  of total size  $|A| + |B| + |C| \leq n$ , solves the underlying 3SUM<sup>+</sup> instance in expected  $\mathcal{O}(n^{2-\varepsilon})$  time. Then, the 3SUM conjecture fails.*

**Proof.** We shall demonstrate an algorithm violating the 3SUM conjecture via Lemma 20. If the input instance already satisfies  $N \leq n^{2+\delta}$ , there is nothing to do. Thus, we henceforth assume  $N > n^{2+\delta}$ . Let  $v = \lceil \log 3N \rceil$  and  $u = \lfloor \log n^{2+\delta} \rfloor$ . In the preprocessing, we draw a uniformly random *odd* integer  $\alpha \in [0..2^v)$ , which defines a hash function  $h : \mathbb{Z} \rightarrow [0..2^u)$  with  $h(x) = \lfloor \frac{\alpha x \bmod 2^v}{2^v - u} \rfloor$  for  $x \in \mathbb{Z}$ . The key property of this function is that  $(h(a) + h(b) + h(c) - h(a + b + c)) \bmod 2^u \in \{0, -1, -2\}$  holds for all  $a, b, c \in \mathbb{Z}$ . At the preprocessing stage, we also preprocess  $(n, 2^u, h(A))$  for the hypothetical 3SUM<sup>+</sup> algorithm (note that  $2^u \leq n^{2+\delta}$ ). Overall, the preprocessing stage costs  $\mathcal{O}(n^d)$  time.

In the main phase, we solve the following 3SUM<sup>+</sup> instances, each of size at most  $n$  and over universe  $[-2^u..2^u)$ , denoting  $X + y := \{x + y : x \in X\}$ :

- $(h(A), h(B), h(C))$ ,
- $(h(A), h(B), h(C) - 2^u + 2)$ ,
- $(h(A), h(B), h(C) - 2^u + 1)$ ,
- $(h(A), h(B), h(C) - 2^u)$ ,
- $(h(A), h(B) - 2^u, h(C) - 2^u + 2)$ ,
- $(h(A), h(B) - 2^u, h(C) - 2^u + 1)$ ,
- $(h(A), h(B) - 2^u, h(C) - 2^u)$ ;

this step costs  $\mathcal{O}(n^{2-\varepsilon})$  time. Combining the results of these calls, in  $\mathcal{O}(n)$  time we derive

$$S := \{c \in C : h(a) + h(b) + h(c) \in \{0, 2^u - 2, 2^u - 1, 2^u, 2 \cdot 2^u - 2, 2 \cdot 2^u - 1, 2 \cdot 2^u\}\}.$$

## 18:12 The Dynamic $k$ -Mismatch Problem

Finally, for each  $c \in S$ , we check in  $\mathcal{O}(n)$  time whether  $a + b + c = 0$  holds for some  $a \in A$  and  $b \in B$ . Upon encountering the first witness  $c \in S$ , we return YES. If no witness is found, we return NO.

Let us analyze the correctness of this reduction. If we return YES, then clearly  $a + b + c = 0$  holds for some  $a \in A$ ,  $b \in B$ , and  $c \in C$ . For the converse implication, suppose that  $a + b + c = 0$  holds for some  $a \in A$ ,  $b \in B$ , and  $c \in C$ . Then,  $(h(a) + h(b) + h(c) - h(a + b + c)) \bmod 2^u = (h(a) + h(b) + h(c)) \bmod 2^u \in \{0, -1, -2\}$ . Given that  $h(a), h(b), h(c) \in [0..2^u)$ , this means that  $h(a) + h(b) + h(c) \in \{0, 2^u - 2, 2^u - 1, 2^u, 2 \cdot 2^u - 2, 2 \cdot 2^u - 1, 2 \cdot 2^u\}$ , i.e.,  $c \in S$ . Consequently, we are guaranteed to return YES while processing  $c \in S$  at the latest.

It remains to bound the expected running time. For this, it suffices to prove that there are, in expectation,  $\mathcal{O}(n^{1-\delta})$  triples  $(a, b, c) \in A \times B \times C$  such that  $a + b + c \neq 0$  yet  $h(a) + h(b) + h(c) \in \{0, 2^u - 2, 2^u - 1, 2^u, 2 \cdot 2^u - 2, 2 \cdot 2^u - 1, 2 \cdot 2^u\}$  (in particular, this means that, in expectation,  $S$  contains at most  $\mathcal{O}(n^{1-\delta})$  non-witnesses; verifying all of them costs  $\mathcal{O}(n^{2-\delta})$  expected time in total). Specifically, we shall prove that each triple satisfies the aforementioned condition with probability  $\mathcal{O}(n^{-2-\delta})$ .

Due to the fact that  $(h(a) + h(b) + h(c) - h(a + b + c)) \bmod 2^u \in \{0, -1, -2\}$ , the bad event holds only if  $a + b + c \neq 0$  yet  $h(a + b + c) \in \{0, 1, 2, 2^u - 2, 2^u - 1\}$ . Let  $a + b + c = 2^t \beta$  for an integer  $t \in \mathbb{Z}_{\geq 0}$  and odd integer  $\beta \in \mathbb{Z}$ . Due to  $|a + b + c| \leq 3N \leq 2^v$ , we must have  $t \in [0..v)$ .

- If  $t > v - u + 1$ , then  $h(a + b + c)$  is uniformly random odd multiple of  $2^{t-v+u}$  within  $[0..2^u)$ . Hence,  $\Pr[h(a + b + c) \in \{0, 1, 2, 2^u - 2, 2^u - 1\}] = 0$ .
- If  $t = v - u + 1$ , then  $h(a + b + c)$  is a uniformly random odd multiple of 2 within  $[0..2^u)$ . Hence,  $\Pr[h(a + b + c) \in \{0, 1, 2, 2^u - 2, 2^u - 1\}] \leq \frac{2}{2^{v-2}} = \frac{8}{2^v}$ .
- If  $t = v - u$ , then  $h(a + b + c)$  is a uniformly random odd multiple of 1 within  $[0..2^u)$ . Hence,  $\Pr[h(a + b + c) \in \{0, 1, 2, 2^u - 2, 2^u - 1\}] \leq \frac{2}{2^{v-1}} = \frac{4}{2^v}$ .
- If  $t < v - u$ , then  $h(a + b + c)$  is a uniformly random element of  $[0..2^u)$ . Hence,  $\Pr[h(a + b + c) \in \{0, 1, 2, 2^u - 2, 2^u - 1\}] \leq \frac{5}{2^v}$ .

Overall, the probability is bounded by  $\frac{8}{2^v} = \mathcal{O}(n^{-2-\delta})$ . ◀

We are now in a position to give the lower bound for the dynamic  $k$ -mismatch problem.

► **Theorem 23.** *Suppose that, for some constants  $p > 0$ ,  $\varepsilon > 0$ , and  $0 < c < \frac{1}{2}$ , there exists a dynamic  $k$ -mismatch algorithm that solves instances satisfying  $k = \lceil m^c \rceil$  using initialization in  $\mathcal{O}(n^p)$  expected time, updates in  $\mathcal{O}(k^{1-\varepsilon})$  expected time, and queries in  $\mathcal{O}(k^{1-\varepsilon})$  expected time. Then, the 3SUM conjecture fails. This statement remains true when updates are allowed in either the pattern or the text (but not both).*

**Proof.** We shall provide an algorithm contradicting Proposition 22 for  $\delta = \frac{1-2c}{2c}$  and  $d = \frac{p}{c}$ . Suppose that the task is to solve a size- $\hat{n}$  instance of the 3SUM<sup>+</sup> problem with  $A, B, C \subseteq [-N..N)$ . We set  $m = \lceil \hat{n}^{1/c} \rceil$  (so that  $k = \lceil m^c \rceil \geq \hat{n}$ ), and  $n = 2m$ , and we initialize a pattern to  $P = 0^m$  and a text to  $T = 0^n$ . Observe that  $m \geq \hat{n}^{1/c} \geq N^{\frac{1}{c(2+\delta)}} = N^{\frac{2}{1+2c}}$ . If  $N^{\frac{2}{1+2c}} < 2N$ , then  $N = \mathcal{O}(1)$ , and we can afford to solve the 3SUM<sup>+</sup> instance naively. Otherwise, we are guaranteed that  $m \geq 2N$ , and we proceed as follows:

- we set  $P[a + N] := 1$  for each  $a \in A$ ;
- we set  $T[2N - b] := 1$  for each  $b \in B$ .

Finally, for each element  $c \in C$ , we perform a query at position  $c + N$ , and report  $c$  if and only if  $\text{HD}(P, T[c + N..c + N + m]) < \text{HD}(P, 0^m) + \text{HD}(T[c + N..c + N + m], 0^m)$ . Due to the fact that  $\text{HD}(P, 0^m) + \text{HD}(T[c + N..c + N + m], 0^m) \leq |A| + |B| \leq \hat{n} = k$ , this can be decided based on the answer to the query. Equivalently, we report  $c \in C$  if and only if

$P[i] = T[c + N + i] = 1$  holds for some  $i \in [0..N)$ , i.e.,  $T[a + c + 2N] = 1$  for some  $a \in A$ , or, equivalently,  $-a - c \in B$ , i.e.,  $a + b + c = 0$  for some  $b \in B$ . This proves the correctness of the algorithm.

As for the running time, note that the preprocessing phase costs  $\mathcal{O}(n^p) = \mathcal{O}(m^p) = \mathcal{O}(\hat{n}^{\frac{p}{c}}) = \mathcal{O}(\hat{n}^d)$  expected time. The main phase, on the other hand, involves  $\mathcal{O}(\hat{n})$  updates and queries, which cost  $\mathcal{O}(\hat{n} \cdot k^{1-\varepsilon}) = \mathcal{O}(\hat{n}^{2-\varepsilon})$  expected time in total. By Proposition 22, this algorithm for 3SUM<sup>+</sup> would violate the 3SUM conjecture.

If the updates are allowed in the text only, we set up the pattern during the preprocessing phase based on the fact that the target value of  $P$  depends on  $A$  only. If the updates are allowed in the pattern only, we exchange the roles of  $A$  and  $B$  and set up the text during the preprocessing phase. ◀

Next, we note that the lower bound can be naturally extended to  $c \geq \frac{1}{2}$ .

► **Corollary 24.** *Suppose that, for some constants  $p > 0$ ,  $\varepsilon > 0$ , and  $0 < c \leq 1$ , there exists a dynamic  $k$ -mismatch algorithm that solves instances satisfying  $k = \lceil m^c \rceil$  using initialization in  $\mathcal{O}(n^p)$  expected time, updates in  $\mathcal{O}(\min(\sqrt{m}, k)^{1-\varepsilon})$  expected time, and queries in  $\mathcal{O}(\min(\sqrt{m}, k)^{1-\varepsilon})$  expected time. Then, the 3SUM conjecture fails. This statement remains true when updates are allowed in either the pattern or the text (but not both).*

**Proof.** When  $c < \frac{1}{2}$ , the result holds directly due to Theorem 23. When  $c \geq \frac{1}{2}$ , we prove that the 3SUM conjecture would be violated through Theorem 23 with  $\hat{c} = \frac{1-\varepsilon}{2-\varepsilon}$  and  $\hat{\varepsilon} = \frac{\varepsilon}{2}$ . Since the  $\hat{k}$ -mismatch problem with  $\hat{k} = \lceil m \rceil^{\hat{c}}$  can be simulated using an instance of the  $k$ -mismatch problem with  $k = \lceil m \rceil^c$ , we note that, in the former setting, the queries and updates can be hypothetically implemented in  $\mathcal{O}((\sqrt{m})^{1-\varepsilon}) = \mathcal{O}(\hat{k}^{\frac{1-\varepsilon}{2-\varepsilon}}) = \mathcal{O}(\hat{k}^{\frac{2-\varepsilon}{2}}) = \mathcal{O}(\hat{k}^{1-\hat{\varepsilon}})$  expected time, violating the 3SUM conjecture via Theorem 23. ◀

## 4.1 Lower Bound for $m \ll n$

While most of the work in this paper focuses on the case where the length of the pattern is linear in the length of the text, for completeness, we provide a lower bound that is only of interest when the pattern is considerably shorter. Our lower bound is conditioned on the Online Matrix-Vector Multiplication conjecture [18], which is often used in the context of dynamic algorithms.

In the Online Boolean Matrix-Vector Multiplication (OMv) problem, we are given as input a Boolean matrix  $M \in \{0, 1\}^{n \times n}$ . Then, a sequence of  $n$  vectors  $v_1, \dots, v_n \in \{0, 1\}^n$  arrives in an online fashion. For each such vector  $v_i$ , we are required to output  $Mv_i$  before receiving  $v_{i+1}$ .

► **Conjecture 25 (OMv Conjecture [18]).** *For any constant  $\epsilon > 0$ , there is no  $\mathcal{O}(n^{3-\epsilon})$ -time algorithm that solves OMv correctly with probability at least  $\frac{2}{3}$ .*

We use the following simplified version of [18, Theorem 2.2].

► **Theorem 26 ([18]).** *Suppose that, for some constants  $\gamma, \varepsilon > 0$ , there is an algorithm that, given as input a matrix  $M \in \{0, 1\}^{p \times q}$ , with  $q = \lfloor p^\gamma \rfloor$ , preprocesses  $M$  in time polynomial in  $p \cdot q$ , and then, presented with a vector  $v \in \{0, 1\}^q$ , computes  $Mv$  in time  $\mathcal{O}(p^{1+\gamma-\varepsilon})$  correctly with probability at least  $\frac{2}{3}$ . Then, the OMv conjecture fails.*

Theorem 26 lets us derive our lower bound for the dynamic  $k$ -mismatch problem.

► **Theorem 27.** *Suppose that, for some constants  $\gamma, \varepsilon > 0$ , there is a dynamic  $k$ -mismatch algorithm that solves instances satisfying  $k = 2 \lfloor (\frac{n}{m})^\gamma \rfloor$ , with preprocessing in  $\mathcal{O}(n^{\mathcal{O}(1)})$  time, updates of the pattern in  $\mathcal{O}((\frac{n}{m})^{1-\varepsilon})$  time, and queries in  $\mathcal{O}(k^{1-\varepsilon})$  time, providing correct answers with high probability. Then, the OMv conjecture fails.*

**Proof.** Given a matrix  $M \in \{0, 1\}^{p \times q}$ , we set  $m = 3q$ ,  $n = 3pq$ , and  $k = 2q$  (so that  $k = 2q = 2 \lfloor p^\gamma \rfloor = 2 \lfloor (\frac{n}{m})^\gamma \rfloor$  holds). At the preprocessing phase, we initially set  $P = 0^m$  and  $T = 0^n$ . As for the text, for each  $i \in [0..p)$  and  $j \in [0..q)$ , we set  $T[3iq + 3j..3iq + 3j + 3)$  to 100 if  $M[i, j] = 0$  and to 111 if  $M[i, j] = 1$ .

When a vector  $v$  arrives, the pattern is set as follows: for each  $j \in [0..q)$ , we set  $P[3j..3j + 3)$  to 001 if  $v[j] = 0$  and to 111 if  $v[j] = 1$ . This requires  $\mathcal{O}(q)$  update calls to our dynamic data structure. Queries are then made at position  $im$  for all  $i \in [0..p)$ . By definition of the Hamming distance,  $\text{HD}(100, 001) = \text{HD}(100, 111) = \text{HD}(111, 001) = 2$ , whereas  $\text{HD}(111, 111) = 0$ ; therefore, the only time that the returned Hamming distance will be less than  $k = 2q$  is when  $M[i, j] = v[j] = 1$  for some  $j \in [1..q]$ , i.e.,  $(Mv)[i] = 1$ . Therefore, the OMv product can be computed by making  $\mathcal{O}(p)$  queries and  $\mathcal{O}(q)$  updates to the dynamic  $k$ -mismatch data structure as before. The total cost of these operations is  $\mathcal{O}(pq^{1-\varepsilon} + qp^{1-\varepsilon}) = \mathcal{O}(p^{1+\gamma(1-\varepsilon)} + p^{1+\gamma-\varepsilon}) = \mathcal{O}(p^{1+\gamma-\min(1,\gamma)\varepsilon})$ . By Theorem 26 with  $\hat{\varepsilon} = \min(1, \gamma)\varepsilon$ , this would violate the OMv conjecture. ◀

---

## References

- 1 Stephen Alstrup, Gerth Stølting Brodal, and Theis Rauhe. Pattern matching in dynamic texts. In *SODA*, pages 819–828, 2000. URL: <http://dl.acm.org/citation.cfm?id=338219.338645>.
- 2 Amihood Amir and Itai Boneh. Update query time trade-off for dynamic suffix arrays. In Yixin Cao, Siu-Wing Cheng, and Minming Li, editors, *31st International Symposium on Algorithms and Computation, ISAAC 2020*, volume 181 of *LIPICs*, pages 63:1–63:16. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.ISAAC.2020.63.
- 3 Amihood Amir and Itai Boneh. Dynamic suffix array with sub-linear update time and poly-logarithmic lookup time. *CoRR*, 2021. arXiv:2112.12678.
- 4 Amihood Amir, Itai Boneh, Panagiotis Charalampopoulos, and Eitan Konradovsky. Repetition detection in a dynamic string. In *ESA*, volume 144 of *LIPICs*, pages 5:1–5:18, 2019. doi:10.4230/LIPICs.ESA.2019.5.
- 5 Amihood Amir, Panagiotis Charalampopoulos, Solon P. Pissis, and Jakub Radoszewski. Longest common substring made fully dynamic. In *ESA*, volume 144 of *LIPICs*, pages 6:1–6:17, 2019. doi:10.4230/LIPICs.ESA.2019.6.
- 6 Amihood Amir, Moshe Lewenstein, and Ely Porat. Faster algorithms for string matching with  $k$  mismatches. *Journal of Algorithms*, 50(2):257–275, 2004. doi:10.1016/S0196-6774(03)00097-X.
- 7 Ilya Baran, Erik D. Demaine, and Mihai Patrascu. Subquadratic algorithms for 3sum. *Algorithmica*, 50(4):584–596, 2008. doi:10.1007/s00453-007-9036-3.
- 8 Timothy M. Chan. More logarithmic-factor speedups for 3SUM, (median, +)-convolution, and some geometric 3SUM-hard problems. *ACM Transactions on Algorithms*, 16(1):7:1–7:23, 2020. doi:10.1145/3363541.
- 9 Timothy M. Chan, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. Approximating text-to-pattern Hamming distances. In *STOC*, pages 643–656. ACM, 2020. doi:10.1145/3357713.3384266.
- 10 Panagiotis Charalampopoulos, Paweł Gawrychowski, and Karol Pokorski. Dynamic longest common substring in polylogarithmic time. In *ICALP*, volume 168 of *LIPICs*, pages 27:1–27:19, 2020. doi:10.4230/LIPICs.ICALP.2020.27.

- 11 Panagiotis Charalampopoulos, Tomasz Kociumaka, and Philip Wellnitz. Faster approximate pattern matching: A unified approach. In *FOCS*, pages 978–989, 2020. doi:10.1109/FOCS46700.2020.00095.
- 12 Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana Starikovskaya. The  $k$ -mismatch problem revisited. In *SODA*, pages 2039–2052. SIAM, 2016. doi:10.1137/1.9781611974331.ch142.
- 13 Raphaël Clifford, Allan Grønlund, Kasper Green Larsen, and Tatiana Starikovskaya. Upper and lower bounds for dynamic data structures on strings. In *STACS*, volume 96 of *LIPIcs*, pages 22:1–22:14, 2018. doi:10.4230/LIPIcs.STACS.2018.22.
- 14 Raphaël Clifford, Tomasz Kociumaka, and Ely Porat. The streaming  $k$ -mismatch problem. In *SODA*, pages 1106–1125. SIAM, 2019. doi:10.1137/1.9781611975482.68.
- 15 Paweł Gawrychowski, Adam Karczmarz, Tomasz Kociumaka, Jakub Łącki, and Piotr Sankowski. Optimal dynamic strings. In *SODA*, pages 1509–1528. SIAM, 2018. doi:10.1137/1.9781611975031.99.
- 16 Paweł Gawrychowski and Przemysław Uznański. Towards unified approximate pattern matching for Hamming and  $L_1$  distance. In *ICALP*, volume 107 of *LIPIcs*, pages 62:1–62:13, 2018. doi:10.4230/LIPIcs.ICALP.2018.62.
- 17 Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. The streaming  $k$ -mismatch problem: Tradeoffs between space and total time. In *CPM*, volume 161 of *LIPIcs*, pages 15:1–15:15, 2020. doi:10.4230/LIPIcs.CPM.2020.15.
- 18 Monika Henzinger, Sebastian Krinninger, Danupon Nanongkai, and Thatchaphol Saranurak. Unifying and strengthening hardness for dynamic problems via the online matrix-vector multiplication conjecture. In *STOC*, pages 21–30. ACM, 2015. doi:10.1145/2746539.2746609.
- 19 Chloe Ching-Yun Hsu and Chris Umans. On multidimensional and monotone  $k$ -sum. In Kim G. Larsen, Hans L. Bodlaender, and Jean-François Raskin, editors, *42nd International Symposium on Mathematical Foundations of Computer Science, MFCS 2017*, volume 83 of *LIPIcs*, pages 50:1–50:13. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.MFCS.2017.50.
- 20 Dominik Kempa and Tomasz Kociumaka. Dynamic suffix array with polylogarithmic queries and updates. *CoRR*, 2022. arXiv:2201.01285.
- 21 Gad M. Landau and Uzi Vishkin. Efficient string matching with  $k$  mismatches. *Theoretical Computer Science*, 43:239–249, 1986. doi:10.1016/0304-3975(86)90178-7.
- 22 Andrea Lincoln, Virginia Vassilevska Williams, Joshua R. Wang, and R. Ryan Williams. Deterministic time-space trade-offs for  $k$ -sum. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016*, volume 55 of *LIPIcs*, pages 58:1–58:14. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPIcs.ICALP.2016.58.
- 23 Mihai Pătraşcu. Towards polynomial lower bounds for dynamic problems. In *STOC*, pages 603–610. ACM, 2010. doi:10.1145/1806689.1806772.