

# Improved Circular $k$ -Mismatch Sketches

Shay Golan 

Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel  
golansh1@cs.biu.ac.il

Tomasz Kociumaka 


Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel  
kociumaka@mimuw.edu.pl

Tsvi Kopelowitz 

Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel  
kopelot@gmail.com

Ely Porat 

Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel  
porately@cs.biu.ac.il

Przemysław Uznański 

Institute of Computer Science, University of Wrocław, Poland  
puznanski@cs.uni.wroc.pl

---

## Abstract

The shift distance  $\text{sh}(S_1, S_2)$  between two strings  $S_1$  and  $S_2$  of the same length is defined as the minimum Hamming distance between  $S_1$  and any rotation (cyclic shift) of  $S_2$ . We study the problem of sketching the shift distance, which is the following communication complexity problem: Strings  $S_1$  and  $S_2$  of length  $n$  are given to two identical players (encoders), who independently compute sketches (summaries)  $\text{sk}(S_1)$  and  $\text{sk}(S_2)$ , respectively, so that upon receiving the two sketches, a third player (decoder) is able to compute (or approximate)  $\text{sh}(S_1, S_2)$  with high probability.

This paper primarily focuses on the more general  $k$ -mismatch version of the problem, where the decoder is allowed to declare a failure if  $\text{sh}(S_1, S_2) > k$ , where  $k$  is a parameter known to all parties. Andoni et al. (STOC'13) introduced exact circular  $k$ -mismatch sketches of size  $\tilde{O}(k + D(n))$ , where  $D(n)$  is the number of divisors of  $n$ . Andoni et al. also showed that their sketch size is optimal in the class of linear homomorphic sketches.

We circumvent this lower bound by designing a (non-linear) exact circular  $k$ -mismatch sketch of size  $\tilde{O}(k)$ ; this size matches communication-complexity lower bounds. We also design  $(1 \pm \varepsilon)$ -approximate circular  $k$ -mismatch sketch of size  $\tilde{O}(\min(\varepsilon^{-2}\sqrt{k}, \varepsilon^{-1.5}\sqrt{n}))$ , which improves upon an  $\tilde{O}(\varepsilon^{-2}\sqrt{n})$ -size sketch of Crouch and McGregor (APPROX'11).

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Pattern matching; Theory of computation  $\rightarrow$  Sketching and sampling

**Keywords and phrases** Hamming distance,  $k$ -mismatch, sketches, rotation, cyclic shift, communication complexity

**Digital Object Identifier** 10.4230/LIPIcs.APPROX/RANDOM.2020.46

**Category** APPROX

**Funding** This work was supported in part by ISF grants no. 1278/16 and 1926/19, by a BSF grant no. 2018364, and by an ERC grant MPM under the EU's Horizon 2020 Research and Innovation Programme (grant no. 683064).

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



© Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, Ely Porat, and Przemysław Uznański; licensed under Creative Commons License CC-BY

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020).

Editors: Jarosław Byrka and Raghu Meka; Article No. 46; pp. 46:1–46:24



Leibniz International Proceedings in Informatics

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

## 1 Introduction

The *Hamming distance* [25] is a fundamental metric for strings, and computing the Hamming distances in various settings is a central task in text processing. The Hamming distance of two length- $n$  strings  $S_1$  and  $S_2$  is defined as the number of aligned mismatches between  $S_1$  and  $S_2$ . In the  $k$ -mismatch variant [1, 4, 14, 22, 32], the problem is parameterized by an integer  $1 \leq k \leq n$ , and the task is relaxed so that if  $\text{Ham}(S_1, S_2) > k$ , then instead of computing  $\text{Ham}(S_1, S_2)$ , the algorithm is only required to report that this is the case, without computing the distance. Since computing the exact Hamming distance, both in the classic version and the  $k$ -mismatch version, is challenging under some efficiency constraints, a large body of research [14, 27, 30, 31] focused on the approximation version of both problems. Formally, in the  $(1 \pm \varepsilon)$ -approximation variant of either problem, the problem is parameterized by  $\varepsilon > 0$ , and whenever the algorithm should report  $\text{Ham}(S_1, S_2)$  in the original problem, in the approximation variant, the algorithm may report a  $(1 \pm \varepsilon)$ -approximation of  $\text{Ham}(S_1, S_2)$ .

**Sketching.** Sketching is one of the settings of sublinear algorithms designed for space-efficient and time-efficient processing of big data, with applications in streaming algorithms, signal processing, network traffic monitoring, and other areas [18, 17, 34]. The task of sketching the Hamming distance boils down to constructing two (randomized) functions  $\text{sk} : \Sigma^n \rightarrow \{0, 1\}^*$  and  $\text{dec} : \{0, 1\}^* \times \{0, 1\}^* \rightarrow \mathbb{N}$  such that  $\text{dec}(\text{sk}(S_1), \text{sk}(S_2)) = \text{Ham}(S_1, S_2)$  holds with high probability<sup>1</sup>. The communication-complexity interpretation of this problem involves three players sharing public randomness: two identical encoders and a decoder. The first encoder receives a string  $S_1$ , while the second encoder receives a string  $S_2$ . Each of the encoders needs to independently summarize its string. The summaries (sketches) are then sent to the decoder, whose task is to retrieve  $\text{Ham}(S_1, S_2)$  based on the summaries alone, without access to  $S_1$  or  $S_2$ . The sketching complexity of Hamming distance, which is the size of the sketch, is well understood: the optimal sketch size is  $\tilde{\Theta}(n)$  for the base variant [37, 40],  $\tilde{\Theta}(k)$  for the  $k$ -mismatch variant [26, 37], and  $\tilde{\Theta}(\varepsilon^{-2})$  for the  $(1 \pm \varepsilon)$ -approximate variants [2, 33, 40].<sup>2</sup> Much less is known about the sketching complexity of edit distance: it is  $\tilde{\Theta}(n)$  for the base variant and  $\tilde{O}(k^8)$  for the  $k$ -error variant [9]. Approximate edit distance sketches with super-constant approximation ratios are also known; see e.g. [11, 35].

**The shift distance.** We consider the *shift distance* [5, 6, 19], which is a cyclic variant of Hamming distance. For two strings  $S_1, S_2 \in \Sigma^n$ , the shift distance is defined as the minimum Hamming distance between  $S_1$  and any cyclic shift (rotation) of  $S_2$ . Formally, if  $\text{cyc}$  is a function cyclically shifting a given string (by one position to the left), then  $\text{sh}(S_1, S_2) = \min\{\text{Ham}(S_1, \text{cyc}^m(S_2)) \mid m \in \mathbb{Z}\}$  is the shift distance between  $S_1$  and  $S_2$ . The research on shift distance for sublinear algorithms is mostly motivated by the observation that the shift distance shares many similarities with the fundamental Hamming distance. At the same time, shift distance inherits some of the challenges exhibited in the edit distance, e.g., in the context of low-dimensional embeddings to  $\ell_1$  [29] and asymmetric query complexity [7].

The first sketching scheme for shift distance, by Andoni et al. [6], allows for  $O(\log^2 n)$ -approximation using sketches of size  $\tilde{O}(1)$ . Crouch and McGregor [19] showed  $(1 \pm \varepsilon)$ -approximate sketches for shift distance that use  $\tilde{O}(\varepsilon^{-2}\sqrt{n})$  space. Andoni et al. [5] designed

<sup>1</sup> An event  $\mathcal{E}$  is said to happen with high probability if  $\Pr[\mathcal{E}] \geq 1 - n^{-\Omega(1)}$ .

<sup>2</sup> Throughout this paper, the  $\tilde{\Theta}(\cdot)$ ,  $\tilde{\Omega}(\cdot)$ , and  $\tilde{O}(\cdot)$  notations suppress  $\log^{O(1)} n$  factors.

exact  $k$ -mismatch circular sketches that use  $\tilde{O}(D(n) + k)$  space, where  $D(n)$  is the number of divisors of  $n$ , which is  $n^{\Theta(1/\log \log n)}$  in the worst case. In [5], it is proven that  $\tilde{\Omega}(D(n))$  is a lower bound for any linear homomorphic sketch for the shift distance  $k$ -mismatch problem.<sup>3</sup>

**Our results.** We consider a (slight) generalization of the problem of sketching the shift distance, where the decoder needs to retrieve  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  for every  $m \in \mathbb{Z}$ . We consider the problem both in the exact setting and in the  $(1 \pm \varepsilon)$ -approximation version.

► **Problem 1.1.** An *exact circular  $k$ -mismatch sketch* ( $k$ -ECS) for  $\Pi \subseteq \Sigma^n$  is a pair of randomized functions<sup>4</sup>  $\mathbf{sk} : \Pi \rightarrow \{0, 1\}^*$  and  $\mathbf{dec} : \{0, 1\}^* \times \{0, 1\}^* \times \mathbb{Z} \rightarrow \mathbb{N}$  such that, for every  $S_1, S_2 \in \Pi$  and  $m \in \mathbb{Z}$ , the following holds with high probability:

- if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m) = \text{Ham}(S_1, \text{cyc}^m(S_2))$ ,
- otherwise,  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m) > k$ .

► **Problem 1.2.** A  $(1 \pm \varepsilon)$ -approximate circular  $k$ -mismatch sketch  $((\varepsilon, k)$ -ACS) for  $\Pi \subseteq \Sigma^n$  is a pair of randomized functions  $\mathbf{sk} : \Pi \rightarrow \{0, 1\}^*$  and  $\mathbf{dec} : \{0, 1\}^* \times \{0, 1\}^* \times \mathbb{Z} \rightarrow \mathbb{R}$  such that, for every  $S_1, S_2 \in \Pi$  and  $m \in \mathbb{Z}$ , the following holds with high probability:

- if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m) \in (1 \pm \varepsilon)\text{Ham}(S_1, \text{cyc}^m(S_2))$ ,
- otherwise,  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m) > (1 - \varepsilon)k$ .

In this paper, a sketch for  $\Pi \subseteq \Sigma^n$  is of size  $s$  if for every  $S \in \Pi$ , we have  $|\mathbf{sk}(S)| \leq s$  with high probability. Our results are stated in the following theorems.

► **Theorem 1.3.** *There exists a  $k$ -ECS sketch for  $\Sigma^n$  of size  $\tilde{O}(k)$ .*

► **Theorem 1.4.** *There exists an  $(\varepsilon, k)$ -ACS sketch for  $\Sigma^n$  of size  $\tilde{O}(\min(\varepsilon^{-2}\sqrt{k}, \varepsilon^{-1.5}\sqrt{n}))$ .*

Notice that Theorem 1.3 circumvents the lower bound of Andoni et al. [5] by using non-linear sketches (however, the sketches are still homomorphic). Moreover, Theorem 1.4 improves upon the  $\tilde{O}(\varepsilon^{-2}\sqrt{n})$  size sketches of Crouch and McGregor [19], and also addresses the more general  $k$ -mismatch variant of the problem.

**Decoding efficiency.** We also discuss the efficiency of evaluating  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m)$  for a given  $m \in \mathbb{Z}$  and the efficiency of evaluating or approximating  $\text{sh}(S_1, S_2)$  based on our sketches. We show that the naive solution of minimizing  $\mathbf{dec}(\mathbf{sk}(S_1), \mathbf{sk}(S_2), m)$  across all  $m \in [n]$  can be sped up significantly. Formally, this yields solutions to the following problems.

► **Problem 1.5.** An *exact  $k$ -mismatch shift distance sketch* ( $k$ -ESDS) for  $\Pi \subseteq \Sigma^n$  is a pair of randomized functions  $\mathbf{sk} : \Pi \rightarrow \{0, 1\}^*$  and  $\mathbf{dec}^{\text{sh}} : \{0, 1\}^* \times \{0, 1\}^* \rightarrow \mathbb{N}$  such that, for every  $S_1, S_2 \in \Pi$ , the following holds with high probability:

- if  $\text{sh}(S_1, S_2) \leq k$ , then  $\mathbf{dec}^{\text{sh}}(\mathbf{sk}(S_1), \mathbf{sk}(S_2)) = \text{sh}(S_1, S_2)$ ,
- otherwise,  $\mathbf{dec}^{\text{sh}}(\mathbf{sk}(S_1), \mathbf{sk}(S_2)) > k$ .

► **Problem 1.6.** A  $(1 \pm \varepsilon)$ -approximate  $k$ -mismatch shift distance sketch  $((\varepsilon, k)$ -ASDS) for  $\Pi \subseteq \Sigma^n$  is a pair of randomized functions  $\mathbf{sk} : \Pi \rightarrow \{0, 1\}^*$  and  $\mathbf{dec}^{\text{sh}} : \{0, 1\}^* \times \{0, 1\}^* \rightarrow \mathbb{R}$  such that, for every  $S_1, S_2 \in \Pi$ , the following holds with high probability:

- if  $\text{sh}(S_1, S_2) \leq k$ , then  $\mathbf{dec}^{\text{sh}}(\mathbf{sk}(S_1), \mathbf{sk}(S_2)) \in (1 \pm \varepsilon)\text{sh}(S_1, S_2)$ ,
- otherwise,  $\mathbf{dec}^{\text{sh}}(\mathbf{sk}(S_1), \mathbf{sk}(S_2)) > (1 - \varepsilon)k$ .

The task of designing efficient algorithms for computing our sketches is left open.

<sup>3</sup> A sketch is *homomorphic* if  $\mathbf{sk}(\text{cyc}(S))$  can be retrieved from  $\mathbf{sk}(S)$  and *linear* if  $\mathbf{sk}$  is a linear mapping.

<sup>4</sup> A randomized function  $f : X \rightarrow Y$  is a random variable whose values are functions from  $X$  to  $Y$ .

**Related work.** A problem closely related to the *circular Hamming distances* problem, asking to determine  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  for all  $0 \leq m < n$ , is the *text-to-pattern Hamming distances* problem, where the input consists of a pattern  $P$  (of length  $m$ ) and a text  $T$  (of length  $n$ ), and the task is to compute the Hamming distances between  $P$  and every length- $m$  substring of  $T$ . A straightforward reduction from the circular Hamming distances problem to the text-to-pattern Hamming distances problem is given by  $P = S_1$  and  $T = S_2 \cdot S_2$ .

In the offline setting, including the exact and approximate  $k$ -mismatch variants, we are not aware of any separation between the two problems. The state-of-the-art exact solution combines an  $\tilde{O}(n\sigma)$ -time solution for small alphabets (of size  $\sigma$ ) [21] with an  $\tilde{O}(n + \frac{nk}{\sqrt{m}})$ -time algorithm [22], which culminates a long line of research [1, 4, 14, 32]. The approximate variant can be solved in  $\tilde{O}(\varepsilon^{-1}n)$  time [30, 31]; these results improve upon [27]. On the other hand, sketches for text-to-pattern Hamming distances need to be much larger than circular sketches: already recovering exact occurrences requires  $\Omega(n - m)$  space [8].

Interestingly, both in the exact and in the approximate setting, the sizes of our circular  $k$ -mismatch sketches coincide with the current upper bounds for space usage in the *streaming  $k$ -mismatch* problem. In that model, the text arrives in a stream, one character at a time, and the goal is to compute, or estimate, after the arrival of each text character, the Hamming distance between  $P$  and the current suffix of  $T$ . The state-of-the-art exact algorithm [15] uses  $\tilde{O}(k)$  space and costs  $\tilde{O}(\sqrt{k})$  time per character, which improves upon [14, 23, 36, 38]. A recent approximate streaming algorithm [12] uses  $\tilde{O}(\min(\varepsilon^{-2}\sqrt{k}, \varepsilon^{-1.5}\sqrt{n}))$  space and costs  $\tilde{O}(\varepsilon^{-3})$  time per character, which improves upon [16, 39].

## 2 Algorithmic Overview and Organization

The central technical contribution of our work is a randomized scheme of selecting positions in a given string  $S \in \Sigma^n$  so that if  $f(S) \subseteq \{1, \dots, n\}$  is the set of selected positions, then the following properties hold:  $|f(S)| = \tilde{O}(k)$  with high probability, the selection is preserved by rotations (the selected positions are shifted along with the underlying characters), and  $|f(S) \cap f(T)| \geq k$  with high probability for every  $T \in \Sigma^n$  such that  $\text{Ham}(S, T) \leq k$ .

Unfortunately, for integer exponents  $\alpha \gg k$ , such a selection of positions is infeasible for strings of the form  $S = Q^\alpha$  (that we call *high powers*), which are fixed points of  $\text{cyc}^{n/\alpha}$ . Moreover, the selection of positions is also infeasible for strings with a relatively small Hamming distance to some high power. Hence, we define the problematic strings to be *pseudo-periodic*, exclude them from the selection scheme, and deal with them separately.

**Sketches for non-pseudo-periodic strings.** In Section 4, we construct sketches for non-pseudo-periodic strings using a selection function  $f$  satisfying the aforementioned properties.

Our  $(\varepsilon, k)$ -ACS sketch stores (non-circular) approximate Hamming distance sketches of  $\text{cyc}^i(S)$  for a random sample of  $\tilde{O}(\sqrt{k})$  positions  $i \in f(S)$ . Given the  $(\varepsilon, k)$ -ACS sketches of two strings  $S_1, S_2$  and a shift value  $m$  such that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , with high probability, there is a shift  $i$  such that the non-circular sketches of both  $\text{cyc}^i(S_1)$  and  $\text{cyc}^{i+m}(S_2)$  are available. The decoder uses these approximate Hamming distance sketches to approximate  $\text{Ham}(\text{cyc}^i(S_1), \text{cyc}^{i+m}(S_2)) = \text{Ham}(S_1, \text{cyc}^m(S_2))$ ; see Section 4.1 for details.

Our  $k$ -ECS sketch, for each position  $i \in f(S)$ , stores a (non-circular) sketch of  $\text{cyc}^i(S)$  capable of retrieving each mismatch with probability  $\Theta(\frac{\log n}{k})$ , but no more than  $O(\log n)$  mismatches in total. Given circular sketches of two strings  $S_1, S_2$  and a shift value  $m$  such that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , with high probability, there are at least  $k$  shifts  $i$  such that the non-circular sketches of both  $\text{cyc}^i(S_1)$  and  $\text{cyc}^{i+m}(S_2)$  are available. Each of these  $k$  pairs of

non-circular sketches yields random mismatches between  $S_1$  and  $\text{cyc}^m(S_2)$ . Consequently, with high probability, each mismatch between  $S_1$  and  $\text{cyc}^m(S_2)$  is reported at least once, which allows for the exact computation of  $\text{Ham}(S_1, \text{cyc}^m(S_2))$ ; see Section 4.2 for details.

**Selection function.** The selection function  $f$  for non-pseudo-periodic strings is constructed in Section 5. Our baseline solution is to sample strings of length  $\frac{n}{\gamma k}$  (for a constant  $\gamma$  fixed in Section 4) with rate  $\tilde{O}(\frac{k}{n})$  and, for each sampled string  $u$ , to add to  $f(S)$  the positions where  $u$  occurs in  $S$ . Unfortunately, since substrings could have much more than  $\gamma k$  occurrences, the variance of  $|f(S)|$  could be rather large, and thus substrings with a large number of occurrences need to be excluded from the sample. This workaround is feasible unless highly periodic regions cover most positions of  $S$ ; see Section 5.1, where the properties of  $f$  are proved using concentration arguments (the Chernoff–Hoeffding bound).

In the complementary case of strings mostly covered by highly periodic regions, we utilize the structure of these regions to deterministically select positions. If there are many disjoint regions, it suffices to select the boundaries of the regions. However, in general we follow a more involved approach inspired by [10, 13]: periodic regions are extended as long as the number of mismatches between the extended region and the period of the region is relatively small compared to the length of the extended region. The positions of these mismatches are also added to  $f(S)$ . Selection of  $f$  in this case is the most technically challenging component of our construction; see Section 5.2 for details.

**Sketches for pseudo-periodic strings.** Each pseudo-periodic string can be assigned to the nearest high power (the *base*) so that two pseudo-periodic strings  $S_1, S_2$  satisfy  $\text{Ham}(S_1, S_2) \leq k$  only if they share the same base. Thus, we first design a 0-mismatch circular sketch (of size  $\tilde{O}(1)$ ) to be used for comparing the bases both in the exact and approximate variants.

Our exact  $k$ -mismatch circular sketch stores the mismatches between the string and its base. Once the decoder verifies that  $S_1$  and  $\text{cyc}^m(S_2)$  share the same base, the mismatches between  $S_1$  and  $\text{cyc}^m(S_2)$  are reconstructed from the mismatches between each of the strings  $S_1, \text{cyc}^m(S_2)$  and their common base. The  $(\varepsilon, k)$ -ACS sketch stores only the mismatches between the string and its base at  $\tilde{O}(\frac{n}{\varepsilon\sqrt{k}})$  sampled positions (so that  $\tilde{O}(\varepsilon^{-1}\sqrt{k})$  mismatches are stored with high probability). Once the decoder verifies that  $S_1$  and  $\text{cyc}^m(S_2)$  share the same base, the mismatches between  $S_1$  and  $\text{cyc}^m(S_2)$  at  $\tilde{O}(\frac{n}{\varepsilon^2 k})$  jointly sampled positions are retrieved to estimate  $\text{Ham}(S_1, \text{cyc}^m(S_2))$ ; see Section 6.

**Organization.** In Sections 4 and 5, we describe the main novel ideas and techniques of this paper, which are used in sketches for strings that are not pseudo-periodic. In Section 6, we provide sketches for pseudo-periodic strings, and in Section 7 we combine the sketches of Section 4 with the sketches of Section 6 in order to prove the main theorems. Notice that these two cases require a slight overlap so that whenever  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , one of the cases accommodates *both*  $S_1$  and  $S_2$ . In Section 7, we also develop another  $(\varepsilon, k)$ -ACS sketch, tailored to approximating large distances. This simple construction improves the size of  $(\varepsilon, k)$ -ACS sketches (for  $k \geq \varepsilon n$ ) from  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$  to  $\tilde{O}(\varepsilon^{-1.5}\sqrt{n})$ . Finally, in Appendix A, we describe efficient decoding algorithms for retrieving the shift distance from the encodings developed for the circular  $k$ -mismatch sketches. A few simple or folklore proofs are deferred from Sections 3–7 to Appendix B.

### 3 Preliminaries

For integers  $\ell \leq r$ , we denote  $[\ell..r] = \{\ell, \ell + 1, \dots, r\}$ . Moreover,  $[n] = [1..n]$ .

A string  $S$  of length  $|S| = n$  is a sequence of characters  $S[1]S[2]\cdots S[n]$  over an alphabet  $\Sigma$ ; in this work, we assume that  $\Sigma = [\sigma]$ . The set of all length- $n$  strings over  $\Sigma$  is denoted by  $\Sigma^n$ . A string  $T$  is a *substring* of a string  $S \in \Sigma^n$  if  $T = S[i]S[i+1]\cdots S[j]$  for  $1 \leq i \leq j \leq n$ . In this case, we denote the occurrence of  $T$  at position  $i$  by  $S[i..j]$ . Such an occurrence is a *fragment* of  $S$ . A fragment  $S[i..j]$  is a *prefix* of  $S$  if  $i = 1$  and a *suffix* of  $S$  if  $j = n$ .

**Hamming distance.** The *Hamming distance*  $\text{Ham}(S, T)$  of two strings  $S, T \in \Sigma^n$  is defined as the number of positions  $i \in [n]$  such that  $S[i] \neq T[i]$ . We denote  $\text{MP}(S, T) = \{i \in [n] \mid S[i] \neq T[i]\}$  to be the set of *mismatch positions* and  $\text{MI}(S, T) = \{(i, S[i], T[i]) \mid i \in [n], S[i] \neq T[i]\}$  to be the underlying *mismatch information*. Note that  $\text{Ham}(S, T) = |\text{MP}(S, T)| = |\text{MI}(S, T)|$ .

For a subset  $A \subseteq [n]$ , we denote  $\text{MI}_A(S, T) = \{(i, a, b) \in \text{MI}(S, T) \mid i \in A\}$  and  $\text{Ham}_A(S, T) = |\text{MI}_A(S, T)|$ . The following result, based on the Chernoff bound and proved in Appendix B, shows that  $\text{Ham}_A(S, T)$  for random  $A$  yields an approximation of  $\text{Ham}(S, T)$ .

► **Lemma 3.1.** *Let  $A$  be a random subset of  $[n]$  with elements chosen independently at rate  $p$ . For  $0 < \varepsilon < 1$ , we have  $\Pr[\text{Ham}_A(S, T) \in (1 \pm \varepsilon)p\text{Ham}(S, T)] \geq 1 - 2 \exp\left(-\frac{p\text{Ham}(S, T)\varepsilon^2}{3}\right)$ .*

The triangle inequality yields  $\text{Ham}(S, U) \leq \text{Ham}(S, T) + \text{Ham}(T, U)$  for  $S, T, U \in \Sigma^n$ . The underlying phenomenon also allows retrieving  $\text{MI}(S, U)$  from  $\text{MI}(S, T)$  and  $\text{MI}(T, U)$ . The following fact is proved in Appendix B.

► **Fact 3.2.** *For every  $S, T, U \in \Sigma^n$  and every  $A \subseteq [n]$ , the mismatch information  $\text{MI}_A(S, U)$  can be retrieved from  $\text{MI}_A(S, T)$  and  $\text{MI}_A(T, U)$  in time  $\tilde{O}(\text{Ham}_A(S, T) + \text{Ham}_A(T, U))$ .*

**Periods.** An integer  $p$  is a *period* of  $S \in \Sigma^*$  if and only if  $S[i] = S[i+p]$  for all  $1 \leq i \leq |S| - p$ . The shortest period of  $S$  is denoted  $\text{per}(S)$ . If  $\text{per}(S) \leq \frac{1}{2}|S|$ , we say that  $S$  is *periodic*.

**Rotations.** For a string  $S = S[1]S[2]\cdots S[n]$ , let  $\text{cyc}(S) = S[2]\cdots S[n]S[1]$ . For  $i \in \mathbb{Z}$ , we denote  $i \circlearrowleft n = ((i - 1) \bmod n) + 1$  so that, for  $i \in [n]$ , the value  $(i - 1) \circlearrowleft n$  is the position of  $S[i]$  in  $\text{cyc}(S)$ .<sup>5</sup> Moreover, for  $M \subseteq \mathbb{Z}$ , we denote  $M \circlearrowleft n = \{i \circlearrowleft n \mid i \in M\}$ . For  $P \subseteq [n]$ , let  $\text{rot}_n(P) = \{(i - 1) \circlearrowleft n \mid i \in P\}$  be the rotated set  $P$ .

The *primitive root* of a string  $S$  is the shortest string  $Q$  such that  $S = Q^\alpha$  for an integer  $\alpha \geq 1$ . The length of the primitive root is denoted by  $\text{root}(S)$ . Notice that  $\text{per}(S) \leq \text{root}(S)$ . Moreover, for every  $m, m' \in \mathbb{Z}$ , we have that  $\text{root}(\text{cyc}^m(S)) = \text{root}(S)$ , and  $\text{cyc}^m(S) = \text{cyc}^{m'}(S)$  if and only if  $\text{root}(S) \mid (m - m')$ .

### 4 Sketches for Non-pseudo-periodic Strings

We say that a string  $S \in \Sigma^n$  is  $(\alpha, \beta)$ -*pseudo-periodic* if there exists a string  $S' \in \Sigma^n$ , called an  $(\alpha, \beta)$ -*base* of  $S$ , such that  $\text{root}(S') \leq \frac{n}{\alpha}$  and  $\text{Ham}(S, S') \leq \beta$ .

► **Observation 4.1.** *If  $S$  is  $(\alpha, \beta)$ -pseudo-periodic with an  $(\alpha, \beta)$ -base  $S'$ , then every rotation  $\text{cyc}^m(S)$  with  $m \in \mathbb{Z}$  is also  $(\alpha, \beta)$ -pseudo-periodic and  $\text{cyc}^m(S')$  is an  $(\alpha, \beta)$ -base of  $\text{cyc}^m(S)$ .*

<sup>5</sup> We introduce the  $\circlearrowleft$  operator because positions in strings are indexed from 1 rather than from 0.



Let  $\mathcal{H}_{n,k}$  be the set of strings in  $\Sigma^n$  that are  $(3\gamma k, \gamma k)$ -pseudo-periodic, where  $\gamma$  is the smallest constant such that  $\gamma \geq 14$  and  $\frac{n}{3\gamma k}$  is an integer. In this section, we present two circular sketches for strings in  $\Sigma^n \setminus \mathcal{H}_{n,k}$ : an  $(\varepsilon, k)$ -ACS sketch and a  $k$ -ECS sketch. Both sketches rely on the following result, proved in Section 5.

► **Theorem 4.2.** *For every two integers  $1 \leq k \leq n$ , there exists a randomized function  $f : \Sigma^n \setminus \mathcal{H}_{n,k} \rightarrow 2^{[n]}$  such that the following holds for every  $S_1, S_2 \in \Sigma^n \setminus \mathcal{H}_{n,k}$ :*

1.  $|f(S_1)| = \tilde{O}(k)$  with high probability,
2.  $f(\text{cyc}(S_1)) = \text{rot}_n(f(S_1))$ ,
3. if  $\text{Ham}(S_1, S_2) \leq k$ , then  $|f(S_1) \cap f(S_2)| \geq k$  with high probability.

## 4.1 An $(\varepsilon, k)$ -ACS Sketch

We start with briefly presenting a useful technical tool, that is, the non-circular version of the approximate sketch. We remark that many variants of this sketch exist, with equivalent space complexity. A short proof is given in Appendix B for the sake of completeness.

► **Theorem 4.3** ( $(1 \pm \varepsilon)$ -approximate sketches, folklore). *There exists a  $(1 \pm \varepsilon)$ -approximate sketch  $\text{sk}_\varepsilon$  such that, given  $\text{sk}_\varepsilon(S_1)$  and  $\text{sk}_\varepsilon(S_2)$  for two strings  $S_1, S_2 \in \Sigma^n$ , one can decode  $\text{Ham}(S_1, S_2)$  with a  $(1 \pm \varepsilon)$ -multiplicative error. The sketches use  $\tilde{O}(\varepsilon^{-2})$  space, the decoding algorithm is correct with high probability and costs  $\tilde{O}(\varepsilon^{-2})$  time.*

Next, we describe our sketching scheme and prove that, together with an appropriate decoding algorithm, it forms an  $(\varepsilon, k)$ -ACS sketch for  $\Sigma^n \setminus \mathcal{H}_{n,k}$ .

► **Construction 4.4.** The encoding function  $\text{circ}_{\varepsilon,k} : \Sigma^n \setminus \mathcal{H}_{n,k} \rightarrow \{0,1\}^*$  is defined as follows:

1. Let  $f : \Sigma^n \setminus \mathcal{H}_{n,k} \rightarrow 2^{[n]}$  be the selection function of Theorem 4.2.
2. Let  $\text{sk}_\varepsilon : \Sigma^n \rightarrow \{0,1\}^*$  be the sketch of Theorem 4.3.
3. Let  $A, B \subseteq [n]$  be two subsets<sup>6</sup> with elements sampled independently with rate  $p = 2\sqrt{\frac{\ln n}{k}}$ .
4. For  $S \in \Sigma^n \setminus \mathcal{H}_{n,k}$ , the encoding  $\text{circ}_{\varepsilon,k}(S)$  stores  $(i, \text{sk}_\varepsilon(\text{cyc}^i(S)))$  for  $i \in f(S) \cap (A \cup B)$ .

► **Proposition 4.5.** *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon,k}$  of Construction 4.4, forms an  $(\varepsilon, k)$ -ACS sketch of  $\Sigma^n \setminus \mathcal{H}_{n,k}$ . The size of this sketch is  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$ , and the decoding algorithm costs  $\tilde{O}(\sqrt{k} + \varepsilon^{-2})$  time with high probability.*

**Proof.** Our decoding procedure iterates over  $i \in f(S_1) \cap A$ . If  $i' := (i + m) \circlearrowleft n \in f(S_2) \cap B$ , the procedure retrieves the sketches  $\text{sk}_\varepsilon(\text{cyc}^i(S_1))$  and  $\text{sk}_\varepsilon(\text{cyc}^{i'}(S_2))$  and recovers a  $(1 + \varepsilon)$ -approximation of  $\text{Ham}(\text{cyc}^i(S_1), \text{cyc}^{i'}(S_2)) = \text{Ham}(S_1, \text{cyc}^m(S_2))$ . Otherwise,  $\infty$  is returned.

We now reason that if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then, with high probability,  $i' \in f(S_2) \cap B$  for some  $i \in f(S_1) \cap A$ . By Theorem 4.2,  $|f(S_1) \cap f(\text{cyc}^m(S_2))| \geq k$ . Thus, for any  $i \in f(S_1) \cap f(\text{cyc}^m(S_2))$ , we have that  $i \in f(S_1) \cap A$  with probability  $p$ . Similarly,  $i' \in f(S_2) \cap B$  with probability  $p$ . Since  $A$  and  $B$  are independent, we have a success probability  $p^2$  for each  $i$  independently. The probability of at least one success is at least  $1 - (1 - p^2)^k \geq 1 - n^{-4}$ .

The decoding time is given by the time needed to compute the intersection of  $f(S_1) \cap A$  and  $\text{rot}_n^m(f(S_2) \cap B)$ , which is  $\tilde{O}(\sqrt{k})$  with high probability, and  $\tilde{O}(\varepsilon^{-2})$  time to decode the distance from a single pair of indices  $i, i'$ , provided that the intersection is not empty. ◀

<sup>6</sup> The sketch would remain valid with one subset only. However, introducing the second subset simplifies the arguments and makes the construction more similar to the counterpart for pseudo-periodic strings.

## 4.2 An $k$ -ECS Sketch

We begin with the following corollary of [37, Theorem 5.1]. The original statement in [37] is given for  $A = [n]$  only, but it can be generalized in a straightforward manner, e.g., by replacing all characters at positions in  $[n] \setminus A$  with a fixed character.

► **Theorem 4.6** (based on [37, Theorem 5.1]). *For every  $k \leq n$  and  $A \subseteq [n]$ , there is a sketch  $\text{sk}_{k,A}$  of size  $\tilde{O}(k)$  such that, given  $\text{sk}_{k,A}(S_1)$  and  $\text{sk}_{k,A}(S_2)$  for two strings  $S_1, S_2 \in \Sigma^n$ :*

- *if  $\text{Ham}_A(S_1, S_2) \leq k$ , then the decoding function returns  $\text{MI}_A(S_1, S_2)$ ;*
- *otherwise, if  $\text{Ham}_A(S_1, S_2) > k$ , the decoding function reports that this is the case.*

*The decoding algorithm is correct with high probability and costs  $\tilde{O}(k)$  time.*

► **Construction 4.7.** The encoding function  $\text{circ}_k : \Sigma^n \setminus \mathcal{H}_{n,k} \rightarrow \{0, 1\}^*$  is defined as follows:

1. Let  $f : \Sigma^n \setminus \mathcal{H}_{n,k} \rightarrow 2^{[n]}$  be the selection function of Theorem 4.2.
2. Let  $A \subseteq [n]$  be a subset with elements sampled independently with rate  $p := \frac{9 \ln n}{k}$ .
3. Denote  $t = \lceil 18 \ln n \rceil$ , and let  $\text{sk}_{t,A} : \Sigma^n \rightarrow \{0, 1\}^*$  be the sketch of Theorem 4.6.
4. For  $S \in \Sigma^n \setminus \mathcal{H}_{n,k}$ , the encoding  $\text{circ}_k(S)$  stores the pairs  $(i, \text{sk}_{t,A}(\text{cyc}^i(S)))$  for  $i \in f(S)$ .

► **Proposition 4.8.** *There exists a decoding function which, together with the encoding  $\text{circ}_k$  of Construction 4.7, forms a  $k$ -ECS sketch of  $\Sigma^n \setminus \mathcal{H}_{n,k}$ . The size of this sketch is  $\tilde{O}(k)$ , and the decoding algorithm costs  $\tilde{O}(k)$  time with high probability.*

**Proof.** The decoding procedure iterates over  $i \in f(S_1) \cap \text{rot}_n^m(f(S_2))$ . If the number of such positions is less than  $k$ , then  $\infty$  is returned. Otherwise, for each  $i \in f(S_1) \cap \text{rot}_n^m(f(S_2))$ , we have  $i' := (i + m) \circlearrowleft n \in f(S_2)$ , and the algorithm runs a decoding procedure for  $\text{sk}_{t,A}(\text{cyc}^i(S_1))$  and  $\text{sk}_{t,A}(\text{cyc}^{i'}(S_2))$ . If any such decoding fails, then  $\infty$  is returned. Otherwise, for each mismatch position  $j$  found, say with  $\text{cyc}^i(S_1)[j] \neq \text{cyc}^{i'}(S_2)[j]$ , the algorithm adds  $(i + j) \circlearrowleft n$  to a set  $M$ , initialized as the empty set. Finally, the size  $|M|$  is returned.

The decoding procedure costs  $\tilde{O}(k)$  time, which is needed both to find all the aligned pairs  $i \in f(S_1), i' \in f(S_2)$  by computing the intersection  $f(S_1) \cap \text{rot}_n^m(f(S_2))$  and to retrieve and gather the mismatches obtained from the aligned pairs (in  $\tilde{O}(t) = \tilde{O}(1)$  time per pair).

**Correctness.** Recall that  $\text{MP}(S_1, \text{cyc}^m(S_2))$  is the set of mismatch positions between  $S_1$  and  $\text{cyc}^m(S_2)$ . First, notice that each  $j \in M$  is a mismatch position between  $S_1$  and  $\text{cyc}^m(S_2)$ , since  $\text{cyc}^i(S_1)[j] \neq \text{cyc}^{i'}(S_2)[j]$  is equivalent to  $S_1[(i + j) \circlearrowleft n] \neq S_2[(i + j + m) \circlearrowleft n]$ . Hence,  $M \subseteq \text{MP}(S_1, \text{cyc}^m(S_2))$  and  $|M| \leq \text{Ham}(S_1, \text{cyc}^m(S_2))$ .

Now, we prove that, with high probability, if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then the algorithm reports  $\text{Ham}(S_1, \text{cyc}^m(S_2))$ , and if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) > k$ , then the algorithm reports a value larger than  $k$ . In the case where  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , we have that  $|f(S_1) \cap \text{rot}_n^m(f(S_2))| \geq k$  with high probability due to Theorem 4.2. Moreover, for every  $i \in f(S_1) \cap \text{rot}_n^m(f(S_2))$ , the expected number of positions in  $A \cap \text{MP}(S_1, \text{cyc}^m(S_2))$  is  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \cdot p \leq k \cdot \frac{9 \ln n}{k} = 9 \ln n$ . Hence, by a Chernoff bound  $\Pr[|A \cap \text{MP}(S_1, \text{cyc}^m(S_2))| > 18 \ln n] \leq \exp(-\frac{9 \ln n}{3}) = n^{-3}$ . Thus, when  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , decoding  $\text{sk}_{t,A}(\text{cyc}^i(S_1))$  and  $\text{sk}_{t,A}(\text{cyc}^{i'}(S_2))$  succeeds for all  $i \in f(S_1) \cap \text{rot}_n^m(f(S_2))$  with high probability.

Conditioned on the event that  $|f(S_1) \cap \text{rot}_n^m(f(S_2))| \geq k$  and the decoding algorithm of  $\text{sk}_{t,A}$  is successful, we now prove that  $|M| = \text{Ham}(S_1, \text{cyc}^m(S_2))$ . For each mismatch  $j \in \text{MP}(S_1, \text{cyc}^m(S_2))$ , there is an independent trial associated with each  $i \in f(S_1) \cap \text{rot}_n^m(f(S_2))$ , which is whether  $((j - i) \circlearrowleft n) \in A$  or not. The trial is successful with probability  $p$ . The probability that at least one of those trials succeeds is at least  $1 - (1 - p)^k \geq 1 - n^{-9}$ . Applying the union bound over all  $j \in \text{MP}(S_1, \text{cyc}^m(S_2))$ , we conclude that  $M = \text{MP}(S_1, \text{cyc}^m(S_2))$  and  $|M| = \text{Ham}(S_1, \text{cyc}^m(S_2))$  with high probability.



If  $\text{Ham}(S_1, \text{cyc}^m(S_2)) > k$ , then the decoding algorithm may return  $\infty$  because of  $|f(S_1) \cap \text{rot}_n^m(f(S_2))| < k$  or due to a decoding failure. If neither of these events happen, the algorithm returns  $|M|$ , which is equal to  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  with high probability (as proved above). Thus, in both cases, a value larger than  $k$  is reported.  $\blacktriangleleft$

## 5 Construction of the Selection Function

For  $S \in \Sigma^n$ , let  $S^* = S \cdot S \cdot S \cdots$  be the infinite string which is the infinite concatenation of  $S$  to itself (for any  $i \in \mathbb{N}$ , we have  $S^*[i] = S[i \circ n]$ ). Let  $\ell = \frac{n}{3\gamma k}$  (recall it is an integer). A position  $i \in [n]$  is called *cubic* if  $u_i = S^*[i..i+3\ell-1]$  has  $\text{per}(u_i) \leq \frac{|u_i|}{3} = \ell$ , i.e., if the cyclic fragment of length  $3\ell$  starting at position  $i$  consists of at least three repetitions of the same factor. Otherwise, position  $i$  is called *non-cubic*. We denote the set of cubic positions in a string  $S$  as  $\mathbf{C}(S)$ , and the set of non-cubic positions as  $\mathbf{N}(S)$ . Notice that  $\mathbf{C}(S) \cup \mathbf{N}(S) = [n]$  and  $\mathbf{C}(S) \cap \mathbf{N}(S) = \emptyset$ .

We present two selection techniques, resulting in functions  $f_n$  and  $f_c$ , designed for strings with many non-cubic positions and for strings with many cubic positions, respectively. Both functions satisfy the first two properties of Theorem 4.2 for any string  $S_1 \in \Sigma^n \setminus \mathcal{H}_{n,k}$ . The functions  $f_n$  and  $f_c$  have the third property of Theorem 4.2 if  $|\mathbf{N}(S_1)| \geq \frac{n}{2}$  and if  $|\mathbf{C}(S_1)| \geq \frac{n}{2}$ , respectively. Thus, the function  $f$  defined through  $f(S) = f_n(S) \cup f_c(S)$  satisfies Theorem 4.2.

### 5.1 Selecting Positions for Strings with Many Non-cubic Positions

Throughout this subsection, let  $h : \Sigma^{3\ell} \rightarrow \{0,1\}$  be a hash function assigning values independently to each  $u \in \Sigma^{3\ell}$  such that  $\Pr[h(u) = 1] = \frac{4k \ln n}{n}$ . For clarity, we omit the explicit dependence on  $h$  in our notation. For  $S \in \Sigma^n$ , define  $f_n(S) = \{i \in \mathbf{N}(S) \mid h(u_i) = 1\}$ .

Our proofs rely on the following multiplicative Chernoff–Hoeffding bound:

► **Proposition 5.1** (Corollary of [20, Theorems 1.10.1 and 1.10.5]). *Let  $X_1, \dots, X_n$  be independent random variables taking values in  $[0, M]$ , let  $X = \sum_{i=1}^n X_i$ , and let  $\mu \geq 0$ .*

(a) *If  $\mu \geq \mathbb{E}[X]$ , then, for every  $\delta > 0$ , we have  $\Pr[X \geq (1 + \delta)\mu] \leq \exp(-\frac{\min(\delta, \delta^2)\mu}{3M})$ .*

(b) *If  $\mu \leq \mathbb{E}[X]$ , then, for every  $0 < \delta < 1$ , we have  $\Pr[X \leq (1 - \delta)\mu] \leq \exp(-\frac{\delta^2\mu}{2M})$ .*

We first prove that  $f_n$  satisfies the first property of Theorem 4.2.

► **Lemma 5.2.** *For every  $S \in \Sigma^n$ , we have  $\Pr[|f_n(S)| < 8k \ln n] \geq 1 - n^{-\Omega(1)}$ .*

**Proof.** For each  $u \in \Sigma^{3\ell}$ , we introduce a random variable  $X_u = |\{i \in f_n(S) \mid u_i = u\}|$ ; notice that  $X_u$  depends only on  $h(u)$ , so the variables  $X_u$  are independent. In order to apply Proposition 5.1 for  $|f_n(S)| = \sum_{u \in \Sigma^{3\ell}} X_u$ , we prove that each  $X_u$  is bounded.

First, note that if  $\text{per}(u) \leq \ell$  or  $h(u) = 0$ , then  $X_u = 0$ . Otherwise, as  $u_i = u = u_{i'}$  for  $i < i' \leq i + 3\ell$  implies  $i' - i \geq \text{per}(u) > \ell$ , we conclude that  $X_u = |\{i \in [n] \mid u_i = u\}| \leq \frac{n}{\ell} = 3\gamma k$ . Now,  $\mathbb{E}[|f_n(S)|] = \sum_{i \in \mathbf{N}(S)} \Pr[h(u_i) = 1] = |\mathbf{N}(S)| \cdot \frac{4k \ln n}{n} \leq 4k \ln n$ , so, by Proposition 5.1(a) with  $\delta = 1$ , we have  $\Pr[|f_n(S)| \geq 8k \ln n] \leq \exp(-\frac{4k \ln n}{3 \cdot 3\gamma k}) = n^{-4/(9\gamma)} = n^{-\Omega(1)}$ .  $\blacktriangleleft$

The following lemma states that  $f_n$  satisfies Property 2 of Theorem 4.2.

► **Lemma 5.3.** *For every  $S \in \Sigma^n$ , we have  $f_n(\text{cyc}(S)) = \text{rot}_n(f_n(S))$ .*

**Proof.** Let  $i \in f_n(\text{cyc}(S))$  and let  $u = (\text{cyc}(S))^*[i..i + \ell - 1] = S^*[i + 1..i + \ell]$ . Since  $i \in f_n(\text{cyc}(S))$ , we have that  $\text{per}(u) > \frac{\ell}{3}$  and  $h(u) = 1$ . Therefore,  $(i + 1) \circ n \in f_n(S)$ , which means that  $i \circ n = i \in \text{rot}_n(f_n(S))$ . Hence,  $f_n(\text{cyc}(S)) \subseteq \text{rot}_n(f_n(S))$ . Symmetrically,  $\text{rot}_n(f_n(S)) \subseteq f_n(\text{cyc}(S))$ . Thus,  $f_n(\text{cyc}(S)) = \text{rot}_n(f_n(S))$ .  $\blacktriangleleft$

Finally, the following lemma states that  $f_n$  satisfies Property 3 of Theorem 4.2.

► **Lemma 5.4.** *Suppose that  $S_1, S_2 \in \Sigma^n$  satisfy  $\text{Ham}(S_1, S_2) \leq k$ . If  $|\mathbf{N}(S_1)| \geq \frac{1}{2}n$ , then  $\Pr[|f_n(S_1) \cap f_n(S_2)| \geq k] \geq 1 - n^{-\Omega(1)}$ .*

**Proof.** For each  $i \in [n]$ , let  $u_i = S_1^*[i..i+3\ell-1]$  and  $v_i = S_2^*[i..i+3\ell-1]$ , and let  $\Lambda = \{i \in \mathbf{N}(S_i) \mid u_i = v_i\}$ . Notice that, for  $i \in [n]$ , we have  $u_i \neq v_i$  if and only if  $\text{MP}(S_1, S_2) \cap ([i..i+3\ell-1] \circlearrowleft n) \neq \emptyset$ . Hence, the number of indices  $i \in [n]$  with  $u_i \neq v_i$  is at most  $|\text{MP}(S_1, S_2)| \cdot 3\ell \leq k \cdot \frac{n}{\gamma k} \leq \frac{n}{\gamma}$ . Since  $|\mathbf{N}(S_1)| \geq \frac{1}{2}n$ , then  $|\Lambda| \geq \frac{1}{2}n - \frac{1}{\gamma}n > \frac{1}{3}n$  due to  $\gamma \geq 6$ . Thus,  $\mathbb{E}[|f_n(S_1) \cap f_n(S_2)|] \geq |\Lambda| \cdot \frac{4k \ln n}{n} \geq \frac{4}{3}k \ln n$ . The rest of the proof follows from Proposition 5.1(b) similarly as Proposition 5.1(a) is applied in the proof of Lemma 5.2. ◀

## 5.2 Selecting Positions for Strings with Many Cubic Positions

Recall that our goal is to design a rotation-invariant mechanism for selecting  $\tilde{O}(k)$  indices so that, given two fairly similar strings, at least  $k$  common indices are selected in both strings. In the selection procedure described in Section 5.1, the decision whether or not to include position  $i$  was based on whether or not  $S^*[i..i+3\ell-1] \in \Pi$  for a certain family  $\Pi \subseteq \Sigma^{3\ell}$ . Then, we argued that  $S_1^*[i..i+3\ell-1] = S_2^*[i..i+3\ell-1] \in \Pi$  for at least  $k$  positions  $i \in [n]$ .

Unfortunately, this strategy might be infeasible if  $\mathbf{C}(S)$  is large, that is, when there is a large number of cubic positions in  $S$ . For example, it could be the case that  $S_1^*[i..i+3\ell-1] \neq S_2^*[i..i+3\ell-1]$  holds for  $3\ell k = \frac{n}{\gamma}$  positions  $i \in [n]$ , and  $S_1^*[i..i+3\ell-1] = S_2^*[i..i+3\ell-1] = \mathbf{a}^{3\ell}$  for the remaining  $n - \frac{n}{\gamma}$  positions  $i \in [n]$ . This may happen even if  $\text{Ham}(S_1, \mathbf{a}^n) = \Omega(\frac{n}{\gamma})$ , i.e., for strings far from being  $(3\gamma k, \gamma k)$ -pseudo-periodic.

We begin with some intuition for the construction of the function  $f_c$ . First, suppose that, for each position  $i \in \mathbf{C}(S)$ , we include in  $f_c(S)$  the smallest  $j > i$  such that  $\text{per}(S^*[i..j]) > \text{per}(S^*[i..i+3\ell-1])$ . In other words,  $f_c(S)$  contains the positions following each maximal cyclic fragment of length at least  $3\ell$  and period at most  $\ell$ . Notice that this construction satisfies Property 2 of Theorem 4.2. Moreover, since each position may belong to at most two such maximal repetitions, the number of positions selected is at most  $\frac{2n}{3\ell} = 2\gamma k$  (so that Property 1 of Theorem 4.2 is satisfied), and a substitution of a single character in  $S$  may remove at most two positions from  $f_c(S)$ . However, if the cubic positions are clustered in few blocks, then this mechanism is not enough to guarantee that Property 3 of Theorem 4.2 is satisfied, i.e., that  $|f_c(S_1) \cap f_c(S_2)| \geq k$  when  $\text{Ham}(S_1, S_2) \leq k$ . Hence, instead of selecting just one position  $j$  for each  $i \in \mathbf{C}(S)$ , several positions are selected using a process inspired by [10] with subsequent improvements in [13]: The fragment  $S^*[i..i+3\ell-1]$  is maximally extended to  $S^*[i..i+\tau_i-1]$  so that the period of  $S^*[i..i+\tau_i-1]$  drops to  $\text{per}(S^*[i..i+3\ell-1])$  after  $\Theta(\frac{k}{n}\tau_i)$  substitutions, and the underlying mismatching positions are added to  $f_c(S)$ .

### 5.2.1 Definition of $f_c$

For any  $i \in \mathbf{C}(S^*)$ , let  $u_i = S^*[i..i+3\ell-1]$ , let  $\rho_i = \text{per}(u_i)$ , and let  $\mu_{S,i} = S^*[i..i+\rho_i-1]$ , which is the string period of  $u_i$ . To avoid clutter in the presentation, we use  $\mu_i = \mu_{S,i}$  when  $S$  is clear from context. Notice that, for  $\tau \geq 2\rho_i$ , the string  $\mu_i^*[1.. \tau]$  is the (unique) string of length  $\tau$  with string period  $\mu_i$ .

We are now ready to formally define the concept of extending (to the right) a cubic fragment starting at position  $i$  for as long as the ratio between the length of the extended fragment and the Hamming distance between the extended fragment and the appropriate prefix of  $\mu_i^*$  is large enough. The length of such a (maximal) extended fragment is defined as

$$\tau_{S,i} = \min \left\{ \tau \mid \tau < \frac{n}{\gamma k} \text{Ham}(S^*[i..i+\tau-1], \mu_i^*[1.. \tau]) \right\}.$$

The following lemma shows that  $\tau_{S,i}$  is well-defined, i.e., that the minimum in the definition of  $\tau_{S,i}$  is taken over a non-empty set. The bound  $\tau_{S,i} \leq 2n$  is also useful later on.

► **Lemma 5.5.** *For every  $S \in \Sigma^n \setminus \mathcal{H}_{n,k}$  and  $i \in \mathcal{C}(S)$ , we have  $\tau_{S,i} \leq 2n$ .*

**Proof.** Let  $i \in \mathcal{C}(S)$  and assume by contradiction that  $\tau_{S,i} > 2n$ . This yields

$$2n \geq \frac{n}{\gamma k} \text{Ham}(S^*[i..i+2n-1], \mu_i^*[1..2n]).$$

Moreover,  $S^*[i..i+n-1] = S^*[i+n..i+2n-1]$ , and so, by the triangle inequality,

$$\begin{aligned} 2\gamma k &\geq \text{Ham}(S^*[i..i+2n-1], \mu_i^*[1..2n]) \\ &= \text{Ham}(S^*[i..i+n-1], \mu_i^*[1..n]) + \text{Ham}(S^*[i+n..i+2n-1], \mu_i^*[n+1..2n]) \\ &= \text{Ham}(S^*[i..i+n-1], \mu_i^*[1..n]) + \text{Ham}(S^*[i..i+n-1], \mu_i^*[n+1..2n]) \\ &\geq \text{Ham}(\mu_i^*[1..n], \mu_i^*[n+1..2n]). \end{aligned}$$

Notice that for any strings  $x, y, z$  (with  $|x| = |y|$ ) and any integer  $m$ , we have  $\text{Ham}(x, y) = \frac{1}{m} \text{Ham}(x^m, y^m)$  and  $\text{Ham}(x, y) \leq \text{Ham}(xz, yz)$ . Thus, due to  $|\mu_i| = \rho_i \leq \ell \leq \frac{n}{3\gamma k}$ , we have

$$\begin{aligned} \text{Ham}(\mu_i, \mu_i^*[n+1..n+\rho_i]) &= \frac{1}{3\gamma k} \text{Ham}(\mu_i^*[1..3\gamma k\rho_i], \mu_i^*[n+1..n+3\gamma k\rho_i]) \\ &\leq \frac{1}{3\gamma k} \text{Ham}(\mu_i^*[1..n], \mu_i^*[n+1..2n]) \leq \frac{2\gamma k}{3\gamma k} < 1. \end{aligned}$$

Consequently,  $\mu_i = \mu_i^*[n+1..n+\rho_i] = \text{cyc}^n(\mu_i)$ , which implies  $\rho_i \mid n$  by primitivity of  $\mu_i$  (recall that  $\mu_i = \text{cyc}^m(\mu_i)$  only for  $\rho_i \mid m$ ). Since  $\tau_{S,i} > n$ , we have  $n \geq \frac{n}{\gamma k} \text{Ham}(S^*[i..i+n-1], \mu_i^*[1..n])$ , that is  $\gamma k \geq \text{Ham}(S^*[i..i+n-1], \mu_i^*[1..n]) = \text{Ham}(S^*[i..i+n-1], \mu_i^{n/\rho_i})$ . Hence,  $S^*[i..i+n-1] \in \mathcal{H}_{n,k}$  so, by Observation 4.1,  $S \in \mathcal{H}_{n,k}$ . ◀

Let  $R_{S,i} = [i..i+\tau_i-1]$  be the positions in the extended fragment, and let  $M_{S,i} = \{j \in R_{S,i} \mid S[j] \neq \mu_i^*[j-i+1]\}$  be the set of positions in  $R_{S,i}$  corresponding to mismatches between  $S^*[i..i+\tau_i-1]$  and  $\mu_i^*[1..\tau_i]$ . To avoid clutter in the presentation, we use  $\tau_i = \tau_{S,i}$ ,  $R_i = R_{S,i}$ , and  $M_i = M_{S,i}$  when  $S$  is clear from context. Define

$$f_c(S) = \bigcup_{i \in \mathcal{C}(S)} (M_i \circlearrowleft n) = \{p \circlearrowleft n \mid p \in M_i, i \in \mathcal{C}(S)\}.$$

## 5.2.2 Properties of $f_c$

**Property 1 of Theorem 4.2.** Our strategy for proving an upper bound on the size of  $f_c(S)$  is to associate each  $i \in \mathcal{C}(S)$  with a carefully defined set  $A_i \subseteq R_i$ . We then select a subset  $\Gamma \subseteq \mathcal{C}(S)$  so that the sets  $A_i$  for  $i \in \Gamma$  are disjoint subsets of  $[1..3n]$  and  $\bigcup_{i \in \Gamma} M_i = \bigcup_{i \in \mathcal{C}(S)} M_i$ . Finally, we show that  $|M_i| = O(\frac{\gamma k}{n}|A_i|)$  for each  $i \in \mathcal{C}(S)$ , and so  $|\bigcup_{i \in \mathcal{C}(S)} M_i| = |\bigcup_{i \in \Gamma} M_i| = O(\sum_{i \in \Gamma} \frac{\gamma k}{n}|A_i|) = O(\gamma k)$ .

For each  $R_i$ , consider the set of indices  $j \in R_i$  such that  $[j, j+2\ell] \cap M_i = \emptyset$ . Formally, let  $A_i = \{j \in R_i \mid [j, j+2\ell] \subseteq R_i \setminus M_i\}$ . The following lemma lets us define  $f_c(S)$  as the union of  $M_i \circlearrowleft n$  for a restricted set of values of  $i$ , with the property of having disjoint sets  $A_i$ .

► **Lemma 5.6.** *Let  $i, i' \in \mathcal{C}(S)$ . If  $i < i'$  and  $A_i \cap A_{i'} \neq \emptyset$ , then  $M_{i'} \subseteq M_i$ .*

The following fact is useful in the proof of Lemma 5.6.

► **Fact 5.7** ([24, Lemma 6]). *Let  $S$  be a periodic string. If  $T$  is a substring of  $S$  of length at least  $2\text{per}(S)$ , then  $\text{per}(S) = \text{per}(T)$ .*

**Proof of Lemma 5.6.** Let  $j \in A_i \cap A_{i'}$ . By definition,  $[j \dots j + 2\ell] \subseteq (R_i \setminus M_i) \cap (R_{i'} \setminus M_{i'})$ . Thus,  $\mu_i^*[1+j-i \dots 2\ell+j-i] = S^*[j \dots j+2\ell-1] = \mu_{i'}^*[1+j-i' \dots 2\ell+j-i']$ . Since  $\rho_i = \text{per}(\mu_i^*) \leq \ell$  and  $\rho_{i'} = \text{per}(\mu_{i'}^*) \leq \ell$ , by Fact 5.7, we have  $\rho_i = \text{per}(\mu_i^*) = \text{per}(\mu_i^*[1+j-i \dots 2\ell+j-i]) = \text{per}(\mu_{i'}^*[1+j-i' \dots 2\ell+j-i']) = \rho_{i'}$ . Therefore,  $\mu_{i'}^*[1 \dots \tau_{i'}] = \mu_i^*[i' - i + 1 \dots i' - i + \tau_{i'}]$  (since the two fragments are extensions of the same periodic string with the same period). Hence, for any  $\tau \leq \tau_{i'}$ , we have  $\text{Ham}(S^*[i' \dots i' + \tau - 1], \mu_{i'}^*[1 \dots \tau]) = \text{Ham}(S^*[i' \dots i' + \tau - 1], \mu_i^*[i' - i + 1 \dots i' - i + \tau])$ .

Since  $\min(A_i \cap A_{i'}) \geq i'$  and  $A_i \subseteq R_i$ , we have that  $\tau_i > i' - i$ . Therefore, for  $\tau = i' - i$ , we have  $i' - i \geq \frac{n}{\gamma k} \text{Ham}(S^*[i \dots i + i' - i - 1], \mu_i^*[1 \dots i' - i]) = \frac{n}{\gamma k} \text{Ham}(S^*[i \dots i' - 1], \mu_i^*[1 \dots i' - i])$ .

Thus, for any  $\tau < i' - i + \tau_{i'}$ , we have

$$\begin{aligned} & \frac{n}{\gamma k} \text{Ham}(S^*[i \dots i + \tau - 1], \mu_i^*[1 \dots \tau]) \\ &= \frac{n}{\gamma k} \text{Ham}(S^*[i \dots i' - 1], \mu_i^*[1 \dots i' - i]) + \frac{n}{\gamma k} \text{Ham}(S^*[i' \dots i + \tau - 1], \mu_i^*[i' - i + 1 \dots \tau]) \\ &\leq i' - i + \frac{n}{\gamma k} \text{Ham}(S^*[i' \dots i' - (i' - i) + \tau - 1], \mu_{i'}^*[1 \dots \tau - (i' - i)]) \\ &\leq i' - i + \tau - (i' - i) = \tau. \end{aligned}$$

Consequently,  $\tau_i \geq i' - i + \tau_{i'}$ , which means that  $R_{i'} \subseteq R_i$ . For a proof that  $M_{i'} \subseteq M_i$ , let us choose  $j' \in M_{i'}$ . By definition,  $S[j'] \neq \mu_{i'}^*[j' - i' + 1] = \mu_i^*[j' - i' + 1 + (i' - i)] = \mu_i^*[j' - i + 1]$ . Hence,  $j' \in M_i$ .  $\blacktriangleleft$

Lemma 5.6 implies that for any two indices  $i < i'$ , if  $A_i \cap A_{i'} \neq \emptyset$ , then  $M_{i'} \subseteq M_i$ , and thus it is enough to consider only the index  $i$  when defining  $f_c(S)$ . Therefore, we define  $\Gamma = \{i' \in \mathcal{C}(S) \mid \forall i < i' : A_i \cap A_{i'} = \emptyset\}$ . Notice that, among  $i \in \Gamma$ , all the sets  $A_i$  are disjoint. Moreover, since for any  $i \in \mathcal{C}(S)$  we have  $A_i \subseteq R_i \subseteq [1 \dots 3n]$  by Lemma 5.5, we have  $\sum_{i \in \Gamma} |A_i| = |\bigcup_{i \in \Gamma} A_i| \leq |[1 \dots 3n]| = 3n$ .

For every  $i \in \mathcal{C}(S)$ , we have  $|A_i| \geq |R_i| - 2\ell|M_i| = |R_i| - \frac{2n}{3\gamma k}|M_i|$ . Furthermore,  $|R_i| - 1 \geq \frac{n}{\gamma k}(|M_i| - 1)$  by definition of  $\tau_i = |R_i|$ . Thus,  $|A_i| > \frac{n}{\gamma k}|M_i| - \frac{n}{\gamma k} - \frac{2n}{3\gamma k}|M_i| = \frac{n}{3\gamma k}(|M_i| - 3)$ . Due to  $[i \dots i + \ell] \subseteq A_i$ , we have  $|A_i| \geq \ell = \frac{n}{3\gamma k}$ , and therefore  $|M_i| < \frac{3\gamma k}{n}|A_i| + 3 \leq \frac{3\gamma k}{n}|A_i| + \frac{9\gamma k}{n}|A_i| = \frac{12\gamma k}{n}|A_i|$ . Hence,  $|f_c(S)| \leq \left| \bigcup_{i \in \mathcal{C}(S)} M_i \right| = \left| \bigcup_{i \in \Gamma} M_i \right| \leq \sum_{i \in \Gamma} |M_i| \leq \sum_{i \in \Gamma} \frac{12\gamma k}{n}|A_i| = \frac{12\gamma k}{n} \sum_{i \in \Gamma} |A_i| \leq 36\gamma k$ .

**Property 2 of Theorem 4.2.** The following lemma states that  $f_c$  satisfies Property 2.

► **Lemma 5.8.** For every  $S \in \Sigma^n$ , we have  $f_c(\text{cyc}(S)) = \text{rot}_n(f_c(S))$ .

**Proof.** Let  $j \in f_c(\text{cyc}(S))$ . There exists  $i \in \mathcal{C}(\text{cyc}(S))$  such that  $j \in M_{\text{cyc}(S), i} \circlearrowleft n$ . Let  $j' \in M_{\text{cyc}(S), i}$  such that  $j = j' \circlearrowleft n$ . We distinguish between two cases: if  $i \in [1 \dots n - 1]$ , then, since  $i \in \mathcal{C}(\text{cyc}(S))$ , we have  $i + 1 \in \mathcal{C}(S)$  and  $\tau_{S, i+1} = \tau_{\text{cyc}(S), i}$ . Therefore,  $j' + 1 \in M_{S, i+1}$  and  $(j' + 1) \circlearrowleft n \in f_c(S)$ . Thus,  $j = (j' + 1 - 1) \circlearrowleft n \in \text{rot}_n(f_c(S))$ . If  $i = n$ , then it must be that  $1 \in \mathcal{C}(S)$  and  $\tau_{S, 1} = \tau_{\text{cyc}(S), n}$ . Therefore,  $j' - n + 1 \in M_{S, 1}$  and  $(j' - n + 1) \circlearrowleft n \in f_c(S)$ . Thus,  $j = (j' - n + 1 - 1) \circlearrowleft n \in \text{rot}_n(f_c(S))$ . The converse inclusion holds symmetrically.  $\blacktriangleleft$

**Property 3 of Theorem 4.2.** We first give a lower bound on  $|f_c(S)|$  in terms of  $|\mathcal{C}(S)|$ .

► **Lemma 5.9.** For every string  $S \in \Sigma^n \setminus \mathcal{H}_{n, k}$ , we have  $|f_c(S)| \geq \frac{\gamma k}{3n} |\mathcal{C}(S)|$ .

**Proof.** First, we shall construct a set  $\Delta \subseteq \mathcal{C}(S)$  such that  $\sum_{i \in \Delta} |R_i| \geq |\mathcal{C}(S)|$  and, for any two distinct indices  $i, i' \in \Delta$ , we have  $R_i \cap R_{i'} = \emptyset$ . We build  $\Delta$  iteratively. We start with  $\Delta = \emptyset$  and, as long as  $\mathcal{C}(S) \not\subseteq \bigcup_{i \in \Delta} R_i$ , we add  $\min(\mathcal{C}(S) \setminus \bigcup_{i \in \Delta} R_i)$  to  $\Delta$ . Let  $i < i'$  be

two indices in  $\Delta$ . When  $i'$  was added to  $\Delta$ , we already had  $i \in \Delta$ . Thus,  $R_i$  ends to the left of  $i'$ , which is the starting point of  $R_{i'}$ . Hence,  $R_i \cap R_{i'} = \emptyset$ . The algorithm terminates when  $C(S) \subseteq \bigcup_{i \in \Delta} R_i$ , so  $|C(S)| \leq |\bigcup_{i \in \Delta} R_i| = \sum_{i \in \Delta} |R_i|$ .

For any  $i \in C(S)$ , we have  $|R_i| = \tau_i < \frac{n}{\gamma k} \text{Ham}(S^*[i \dots i + \tau_i - 1], \mu_i^*[1 \dots \tau_i])$ , i.e.,  $|R_i| < \frac{n}{\gamma k} |M_i|$ . Since  $M_i \subseteq R_i$  for every  $i$ , the sets  $M_i$  for  $i \in \Delta$  are disjoint. Consequently,  $|\bigcup_{i \in \Delta} M_i| = \sum_{i \in \Delta} |M_i| > \frac{\gamma k}{n} \sum_{i \in \Delta} |R_i| \geq \frac{\gamma k}{n} |C(S)|$ .

By Lemma 5.5, for any  $i \in C(S)$ , we have  $\tau_i \leq 2n$ . Therefore,  $\bigcup_{i \in \Delta} M_i \subseteq [1 \dots 3n]$  and each position in  $j \in \bigcup_{i \in \Delta} (M_i \circlearrowleft n)$  may be introduced by at most 3 positions  $j, j+n, j+2n \in \bigcup_{i \in \Delta} M_i$ . Thus,  $|f_c(S)| = |\bigcup_{i \in \Delta} (M_i \circlearrowleft n)| \geq \frac{1}{3} |\bigcup_{i \in \Delta} M_i| \geq \frac{\gamma k}{3n} |C(S)|$ . ◀

Using Lemma 5.9, we prove the third property of Theorem 4.2, assuming  $|C(S_1)| \geq \frac{1}{2}n$ .

► **Lemma 5.10.** *Suppose that  $S_1, S_2 \in \Sigma^n \setminus \mathcal{H}_{n,k}$  satisfy  $\text{Ham}(S_1, S_2) \leq k$ . If  $|C(S_1)| \geq \frac{1}{2}n$ , then  $|f_c(S_1) \cap f_c(S_2)| \geq k$ .*

**Proof.** Let  $S'$  be a string of length  $n$ , where, for any  $i$  with  $S_1[i] = S_2[i]$ , we have  $S'[i] = S_1[i]$  and, for any other  $i$  (i.e., for  $i \in \text{MP}(S_1, S_2)$ ), we have  $S'[i] = \$_i$ , where  $\$_i \notin \Sigma$  differs from any other character  $\$_{i'}$  for  $i' \neq i$ .

▷ **Claim 5.11.**  $f_c(S') \subseteq (f_c(S_1) \cap f_c(S_2)) \cup \text{MP}(S_1, S_2)$ .

**Proof.** Let  $j \in f_c(S')$ . If  $j \in \text{MP}(S_1, S_2)$ , the claim follows; thus, assume  $j \notin \text{MP}(S_1, S_2)$ . By the definition of  $f_c(S')$ , there is an index  $i \in C(S')$  such that  $j \in M_{S',i} \circlearrowleft n$ ; let  $j' \in M_{S',i}$  be an integer such that  $j = j' \circlearrowleft n$ . Notice that  $\mu_{S_1,i} = \mu_{S',i}$  since if  $\mu_{S',i}$  contains some  $\$_k$  character, then  $i$  cannot be cubic and so  $i \notin C(S')$ . Therefore,  $\mu_{S_1,i} = \mu_{S',i}$ , and let  $\mu_i = \mu_{S_1,i}$ . For any integer  $\tau$ , we have  $\text{Ham}(S_1^*[i \dots i + \tau - 1], \mu_i^*[1 \dots \tau]) \leq \text{Ham}((S')^*[i \dots i + \tau - 1], \mu_i^*[1 \dots \tau])$  because the new  $\$_k$  characters in  $S'$  just form new mismatches. In particular, for  $\tau_{S_1,i}$  we have  $\frac{n}{\gamma k} \text{Ham}((S')^*[i \dots i + \tau_{S_1,i} - 1], \mu_i^*[1 \dots \tau_{S_1,i}]) \geq \frac{n}{\gamma k} \text{Ham}(S_1^*[i \dots i + \tau_{S_1,i} - 1], \mu_i^*[1 \dots \tau_{S_1,i}]) > \tau_{S_1,i}$ . Hence,  $\tau_{S',i} \leq \tau_{S_1,i}$  and  $R_{S',i} \subseteq R_{S_1,i}$ . Since  $j' \in M_{S',i}$  and  $j \notin \text{MP}(S_1, S_2)$ , it must be that  $j' \in M_{S_1,i}$ . Similarly,  $j' \in M_{S_2,i}$ . Thus,  $j = j' \circlearrowleft n \in (f_c(S_1) \cap f_c(S_2)) \cup \text{MP}(S_1, S_2)$ . ◀

▷ **Claim 5.12.**  $|C(S')| \geq \frac{\gamma-2}{2\gamma}n$ .

**Proof.** Recall that  $|C(S_1)| \geq \frac{1}{2}n$ . If  $\mu_{S_1,i} = \mu_{S',i}$  and  $i \in C(S_1)$ , then  $i \in C(S')$ . The only indices  $i \in C(S_1) \cap \mathbf{N}(S')$  are indices such that  $\mu_{S_1,i} \neq \mu_{S',i}$ , which means that  $\text{MP}(S_1, S_2) \cap ([i \dots i + 3\ell - 1] \circlearrowleft n) \neq \emptyset$ . Hence, each  $m \in \text{MP}(S_1, S_2)$  will remove at most  $3\ell$  positions from  $C(S_1)$ . Thus,  $|C(S')| \geq \frac{1}{2}n - |\text{MP}(S_1, S_2)|3\ell \geq \frac{1}{2}n - k \frac{n}{\gamma k} = \frac{\gamma-2}{2\gamma}n$ . ◀

Due to Claim 5.12, we have  $|C(S')| \geq \frac{\gamma-2}{2\gamma}n$ , and therefore  $|f_c(S')| > \frac{\gamma k}{3n} \frac{\gamma-2}{2\gamma}n = \frac{(\gamma-2)k}{6}$  by Lemma 5.9. Due to Claim 5.11,  $f_c(S') \subseteq (f_c(S_1) \cap f_c(S_2)) \cup \text{MP}(S_1, S_2)$ , and therefore  $|f_c(S')| \leq |(f_c(S_1) \cap f_c(S_2)) \cup \text{MP}(S_1, S_2)| \leq |f_c(S_1) \cap f_c(S_2)| + |\text{MP}(S_1, S_2)| \leq |f_c(S_1) \cap f_c(S_2)| + k$ . Consequently, since  $\gamma \geq 14$ , we have  $|f_c(S_1) \cap f_c(S_2)| \geq \frac{\gamma-8}{6}k \geq \frac{14-8}{6}k = k$ . ◀

## 6 Sketches for Pseudo-periodic Strings

Let  $\mathcal{H}'_{n,k} \subseteq \Sigma^n$  be the family of  $(3\gamma k, (\gamma+1)k)$ -pseudo-periodic strings in  $\Sigma^n$ . In this section, we develop circular sketches for  $\mathcal{H}'_{n,k}$ . We start with a few properties of pseudo-periodic strings. Recall that a string  $S \in \Sigma^n$  is called  $(\alpha, \beta)$ -pseudo-periodic if it has an  $(\alpha, \beta)$ -base  $S' \in \Sigma^n$  with  $\text{root}(S') \leq \frac{n}{\alpha}$  and  $\text{Ham}(S, S') \leq \beta$ . If  $\lfloor \alpha \rfloor > 2\beta$ , then the  $(\alpha, \beta)$ -base is unique.

► **Lemma 6.1.** *If  $S \in \Sigma^n$  is an  $(\alpha, \beta)$ -pseudo-periodic string for some parameters  $\lfloor \alpha \rfloor > 2\beta$ , then  $S'$  has a unique  $(\alpha, \beta)$ -base.*

**Proof.** Suppose that  $S$  has two bases  $S', S''$ . Alzamel et al. [3] show that if  $|X| = |Y| \geq \text{per}(X) + \text{per}(Y)$  and  $X \neq Y$ , then  $\text{Ham}(X, Y) \geq \lfloor \frac{2n}{\text{per}(X) + \text{per}(Y)} \rfloor$ . Setting  $X = S'$  and  $Y = S''$ , we get a contradiction:  $\text{Ham}(S', S'') \geq \lfloor \frac{2n}{\text{per}(S') + \text{per}(S'')} \rfloor \geq \lfloor \frac{2n}{\text{root}(S') + \text{root}(S'')} \rfloor \geq \lfloor \frac{2n}{n/\alpha + n/\alpha} \rfloor = \lfloor \alpha \rfloor > 2\beta \geq \text{Ham}(S, S') + \text{Ham}(S, S'') \geq \text{Ham}(S', S'')$ . ◀

Moreover, the triangle inequality immediately yields the following observation.

► **Observation 6.2.** *Let  $S \in \Sigma^n$  be an  $(\alpha, \beta)$ -pseudo-periodic string and let  $T \in \Sigma^n$  be a string such that  $\text{Ham}(S, T) \leq k$ . Then,  $T$  is  $(\alpha, \beta + k)$ -pseudo-periodic, and every  $(\alpha, \beta)$ -base of  $S$  is an  $(\alpha, \beta + k)$ -base of  $T$ .*

Combining Lemma 6.1 with Observations 4.1 and 6.2, we obtain the following corollary.

► **Corollary 6.3.** *Let  $S_1, S_2 \in \mathcal{H}'_{n,k}$  with  $(3\gamma k, (\gamma + 1)k)$ -bases  $S'_1$  and  $S'_2$ , respectively. If, for some  $m \in \mathbb{Z}$ , we have  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $S'_1 = \text{cyc}^m(S'_2)$ .*

**Proof.** By Observation 6.2,  $S'_1$  is a  $(3\gamma k, (\gamma + 2)k)$ -base of  $\text{cyc}^m(S_2)$ . Moreover, by Observation 4.1,  $\text{cyc}^m(S'_2)$  is a  $(3\gamma k, (\gamma + 1)k)$ -base of  $\text{cyc}^m(S_2)$ , and thus also a  $(3\gamma k, (\gamma + 2)k)$ -base of  $\text{cyc}^m(S_2)$ . Since  $\lfloor 3\gamma k \rfloor > 2(\gamma + 2)k$  due to  $\gamma \geq 5$ , Lemma 6.1 implies that  $S'_1 = \text{cyc}^m(S'_2)$ . ◀

## 6.1 A 0-mismatch Circular Sketch

Both the exact and the  $(1 \pm \varepsilon)$ -approximation sketches of strings in  $\mathcal{H}'_{n,k}$  rely on 0-mismatch circular sketches, which we implement using Karp–Rabin fingerprints.

► **Fact 6.4** (Karp–Rabin fingerprints [28]). *For every positive integer  $n$ , there exists a randomized function  $\Phi : \Sigma^n \rightarrow \{0, 1\}^{O(\log n)}$  such that, for every  $S_1, S_2 \in \Sigma^n$ , the following holds with high probability: if  $S_1 \neq S_2$ , then  $\Phi(S_1) \neq \Phi(S_2)$ .*

**Proof.** The function  $\Phi$  is based on a fixed prime number  $p \geq \max(\sigma, n^2)$  and a uniformly random  $x \in [0 \dots p - 1]$ . The function  $\Phi$  maps a string  $S$  to  $(\sum_{i=1}^{|S|} x^{i-1} \cdot S[i]) \bmod p$ . This way, for every two strings  $S_1 \neq S_2$  in  $\Sigma^n$ , we have  $\Pr[\Phi(S_1) = \Phi(S_2)] \leq \frac{n}{p} \leq \frac{n}{n^2} = n^{-1}$ . ◀

► **Lemma 6.5.** *There exists a 0-ECS sketch  $(\text{sk}_0, \text{dec}_0)$  for  $\Sigma^n$  of size  $O(\log n)$  bits with constant decoding time.*

**Proof.** The construction relies on a Karp–Rabin fingerprint function  $\Phi$ . The sketch  $\text{sk}_0(S)$  for a string  $S \in \Sigma^n$  is defined based on the minimum cyclic rotation of  $S$ , denoted  $\text{minrot}(S)$ , and consists of the following components:

- the fingerprint  $\Phi(\text{minrot}(S))$  of the minimum cyclic rotation of  $S$ ,
- the length  $\text{root}(S)$  of the primitive root of  $S$ ,
- the smallest integer  $r \geq 0$  such that  $S = \text{cyc}^r(\text{minrot}(S))$ .

The decoding function  $\text{dec}_0$  is given two sketches  $\text{sk}_0(S_1) = (\Phi(\text{minrot}(S_1)), \text{root}(S_1), r_1)$ ,  $\text{sk}_0(S_2) = (\Phi(\text{minrot}(S_2)), \text{root}(S_2), r_2)$ , and a shift  $m$ . If  $\Phi(\text{minrot}(S_1)) \neq \Phi(\text{minrot}(S_2))$ , then  $S_1 \neq \text{cyc}^m(S_2)$ , and thus the function returns  $\infty$ . Otherwise,  $\text{minrot}(S_1) = \text{minrot}(S_2)$  with high probability, and the implementation proceeds assuming that  $\text{minrot}(S_1) = T = \text{minrot}(S_2)$  for a string  $T \in \Sigma^n$ . In particular, this implies  $\text{root}(S_1) = \text{root}(T) = \text{root}(S_2)$ . Finally, since  $S_1 = \text{cyc}^{r_1}(T)$  equals  $\text{cyc}^m(S_2) = \text{cyc}^{m+r_2}(T)$  if and only if  $\text{root}(T) \mid (m + r_2 - r_1)$ , the function returns 0 or  $\infty$  depending on whether  $\text{root}(S_1) \mid (m + r_2 - r_1)$  or not. ◀



## 6.2 A $k$ -ECS Sketch

► **Construction 6.6.** The encoding function  $\text{circ}_k : \mathcal{H}'_{n,k} \rightarrow \{0,1\}^*$  is defined as follows:

1. Let  $\text{sk}_0$  be the 0-mismatch sketch of Lemma 6.5.
2. For  $S \in \mathcal{H}'_{n,k}$ , the encoding  $\text{circ}_k(S)$  stores the sketch  $\text{sk}_0(S')$  of the  $(3\gamma k, (\gamma+1)k)$ -base  $S'$  of  $S$  and the mismatch information  $\text{MI}(S, S')$ .

► **Proposition 6.7.** *There exists a decoding function which, together with the encoding  $\text{circ}_k$  of Construction 6.6, forms a  $k$ -ECS sketch of  $\mathcal{H}'_{n,k}$ . The size of the sketch is  $\tilde{O}(k)$ , and the decoding time is  $\tilde{O}(k)$  with high probability.*

**Proof.** The decoding function is given two sketches  $\text{circ}_k(S_1) = (\text{sk}_0(S'_1), \text{MI}(S_1, S'_1))$ ,  $\text{circ}_k(S_2) = (\text{sk}_0(S'_2), \text{MI}(S_2, S'_2))$ , and a shift  $m$ . By Corollary 6.3, if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $S'_1 = \text{cyc}^m(S'_2)$ , and this condition is checked by applying  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$ . If the call returns a non-zero result, then  $\infty$  is returned. Otherwise,  $S'_1 = \text{cyc}^m(S'_2)$  holds with high probability. The analysis below is conditioned on this event.

First,  $\text{MI}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2))$  is retrieved from  $\text{MI}(S_2, S'_2)$  by shifting all the mismatches. Next, the decoding function retrieves  $\text{MI}(S_1, \text{cyc}^m(S_2))$  from  $\text{MI}(S_1, S'_1)$  and  $\text{MI}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2))$  (using Fact 3.2 and assuming that  $S'_1 = \text{cyc}^m(S'_2)$ ) and returns  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = |\text{MI}(S_1, \text{cyc}^m(S_2))|$ . ◀

## 6.3 An $(\varepsilon, k)$ -ACS Sketch

For the pseudo-periodic  $(\varepsilon, k)$ -ACS sketches, we relax the problem statement; we overcome this relaxation in Section 7. In the *relaxed*  $(\varepsilon, k)$ -ACS sketch, the distances smaller than  $\frac{k}{2}$  do not need to be approximated. More precisely, we require the following:

- if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) < \frac{1}{2}k$ , then  $\text{dec}(\text{sk}(S_1), \text{sk}(S_2), m) < \frac{1+\varepsilon}{2}k$ ,
- if  $\frac{1}{2}k \leq \text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $\text{dec}(\text{sk}(S_1), \text{sk}(S_2), m) \in (1 \pm \varepsilon)\text{Ham}(S_1, \text{cyc}^m(S_2))$ ,
- otherwise,  $\text{dec}(\text{sk}(S_1), \text{sk}(S_2), m) > (1 - \varepsilon)k$ .

► **Construction 6.8.** The encoding function  $\text{circ}_{\varepsilon,k} : \mathcal{H}'_{n,k} \rightarrow \{0,1\}^*$  is defined as follows:

1. Let  $\text{sk}_0$  be the 0-mismatch sketch of Lemma 6.5.
2. Let  $A, B \subseteq [n]$  be two subsets with elements sampled independently with rate  $p := \sqrt{\frac{\log n}{\varepsilon^2 k}}$ .
3. For  $S \in \mathcal{H}'_{n,k}$ , the encoding  $\text{circ}_{\varepsilon,k}(S)$  stores the sketch  $\text{sk}_0(S')$  of the  $(3\gamma k, (\gamma+1)k)$ -base  $S'$  of  $S$  and the mismatch information  $\text{MI}_{A \cup B}(S, S')$ .

► **Proposition 6.9.** *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon,k}$  of Construction 6.8, forms a relaxed  $(\varepsilon, k)$ -ACS sketch of  $\mathcal{H}'_{n,k}$ . The size of the sketch is  $\tilde{O}(\varepsilon^{-1}\sqrt{k})$ , and the decoding time is  $\tilde{O}(\varepsilon^{-1}\sqrt{k})$  with high probability.*

**Proof.** The decoding function is given two sketches  $\text{circ}_{\varepsilon,k}(S_1) = (\text{sk}_0(S'_1), \text{MI}_{A \cup B}(S_1, S'_1))$  and  $\text{circ}_{\varepsilon,k}(S_2) = (\text{sk}_0(S'_2), \text{MI}_{A \cup B}(S_2, S'_2))$ , and a shift  $m$ . According to Corollary 6.3, if  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ , then  $S'_1 = \text{cyc}^m(S'_2)$ , and this condition is checked by applying  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$ . If the call returns a non-zero result, then  $\infty$  is returned. Otherwise,  $S'_1 = \text{cyc}^m(S'_2)$  holds with high probability. The analysis below is conditioned on this event.

First,  $\text{MI}_{A \cap \text{rot}_n^m(B)}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2))$  is retrieved by filtering and shifting  $\text{MI}_{A \cup B}(S_2, S'_2)$ . Secondly,  $\text{MI}_{A \cap \text{rot}_n^m(B)}(S_1, S'_1)$  is retrieved by filtering  $\text{MI}_{A \cup B}(S_1, S'_1)$ . Then, the algorithm retrieves  $\text{MI}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  combining  $\text{MI}_{A \cap \text{rot}_n^m(B)}(S_1, S'_1)$  and  $\text{MI}_{A \cap \text{rot}_n^m(B)}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2))$  (using Fact 3.2 and assuming that  $S'_1 = \text{cyc}^m(S'_2)$ ). Since  $A \cap \text{rot}_n^m(B)$  is a random subset of  $[n]$  with elements sampled independently with rate  $\frac{\log n}{\varepsilon^2 k}$ , the quantity  $\frac{\varepsilon^2 k}{\log n} \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  is a  $(1 \pm \varepsilon)$ -approximation of  $\text{Ham}(S_1, S_2)$  with high probability provided that  $\text{Ham}(S_1, S_2) = \Omega(k)$ ; see Lemma 3.1. ◀

## 7 Proofs of Main Theorems

In this section, we complete our construction of circular  $k$ -mismatch sketches for  $\Sigma^n$ .

► **Theorem 1.3.** *There exists a  $k$ -ECS sketch for  $\Sigma^n$  of size  $\tilde{O}(k)$ .*

**Proof.** Our construction combines the  $k$ -ECS sketches of Propositions 6.7 and 4.8. For each string  $S \in \Sigma$ , if  $S \in \mathcal{H}'_{n,k}$ , then the sketch contains the sketch of Proposition 6.7, and if  $S \in \Sigma^n \setminus \mathcal{H}_{n,k}$ , then the sketch contains the sketch of Proposition 4.8. Notice that the sketch contains both components if  $S \in \mathcal{H}'_{n,k} \setminus \mathcal{H}_{n,k}$ .

For two strings  $S_1, S_2 \in \Sigma^n$ , given the sketches of  $S_1$  and  $S_2$ , the decoder works as follows. If the two sketches contain compatible components (of Proposition 4.8 or of Proposition 6.7), then the decoder uses the decoder corresponding to these components. Otherwise, without loss of generality, it must be that  $S_1 \in \mathcal{H}_{n,k}$  and  $S_2 \notin \mathcal{H}'_{n,k}$ . Thus, by Observation 6.2,  $\text{Ham}(S_1, S_2) > k$ , and therefore the decoder outputs  $\infty$ . The decoding time is  $\tilde{O}(k)$ . ◀

Similarly, combining the results of Sections 4 and 6 gives  $(1 + \varepsilon)$ -approximate sketches. The proof of the following result mimics the proof of Theorem 1.3 and is given in Appendix B for completeness.

► **Proposition 7.1.** *There exists a relaxed  $(\varepsilon, k)$ -ACS sketch for  $\Sigma^n$  of size  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$ . Its decoding time is  $\tilde{O}(\varepsilon^{-1}\sqrt{k} + \varepsilon^{-2})$  with high probability.*

A simple alternative approach yields smaller sketches when  $k$  is large compared to  $n$ .

► **Construction 7.2.** The encoding function  $\text{circ}_{\varepsilon,k} : \Sigma^n \rightarrow \{0,1\}^*$  is defined as follows:

1. Let  $A, B \subseteq [n]$  be two subsets with elements sampled independently with rate  $p := \sqrt{\frac{\log n}{\varepsilon^2 k}}$ .
2. For  $S \in \Sigma^n$ , the encoding  $\text{circ}_{\varepsilon,k}(S)$  consists of pairs  $(i, S[i])$  for  $i \in A \cup B$ .

► **Proposition 7.3.** *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon,k}$  of Construction 7.2, forms a relaxed  $(\varepsilon, k)$ -ACS sketch of  $\Sigma^n$ . The size of the sketch is  $\tilde{O}(\frac{n}{\varepsilon\sqrt{k}})$ , and the decoding time is  $\tilde{O}(\frac{n}{\varepsilon\sqrt{k}})$  with high probability.*

**Proof.** The decoder, given the sketches of  $S_1, S_2 \in \Sigma^n$  and a shift  $m$ , uses Lemma 3.1 to estimate  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  based on  $\text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$ . For each  $i \in A \cap \text{rot}_n^m(B)$ , the decoder retrieves  $S_1[i]$  from the sketch of  $S_1$  and  $\text{cyc}^m(S_2)[i] = S_2[(i-m) \circlearrowleft n]$  from the sketch of  $S_2$ . Since  $A \cap \text{rot}_n^m(B)$  is a random subset of  $[n]$  with elements sampled independently with rate  $\frac{\log n}{\varepsilon^2 k}$ , the quantity  $\frac{\varepsilon^2 k}{\log n} \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  is a  $(1 \pm \varepsilon)$ -approximation of  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  with high probability provided that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \Omega(k)$ ; see Lemma 3.1. ◀

► **Theorem 1.4.** *There exists an  $(\varepsilon, k)$ -ACS sketch for  $\Sigma^n$  of size  $\tilde{O}(\min(\varepsilon^{-2}\sqrt{k}, \varepsilon^{-1.5}\sqrt{n}))$ .*

**Proof.** An  $(\varepsilon, k)$ -ACS sketch is obtained by combining  $O(\log k)$  relaxed  $(\varepsilon, k')$ -ACS sketches, where  $k'$  ranges over powers of two between 1 and  $2k$ . Depending on whether  $k' \leq \varepsilon n$  or not, Proposition 7.1 or Proposition 7.3 is used to implement  $k'$ -mismatch sketches. ◀

► **Remark 7.4.** Applying Proposition 7.3 instead of Proposition 7.1 improves the sketch size (for  $k \geq \varepsilon n$ ) but degrades the decoding time. We get two alternatives:  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$ -size sketches with decoding time  $\tilde{O}(\varepsilon^{-1}\sqrt{k} + \varepsilon^{-2})$ , and  $\tilde{O}(\varepsilon^{-1.5}\sqrt{n})$ -size sketches with decoding time  $\tilde{O}(\varepsilon^{-1.5}\sqrt{n})$ .

## References

- 1 Karl R. Abrahamson. Generalized string matching. *SIAM Journal on Computing*, 16(6):1039–1051, 1987. doi:10.1137/0216067.
- 2 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. doi:10.1006/jcss.1997.1545.
- 3 Mai Alzamel, Maxime Crochemore, Costas S. Iliopoulos, Tomasz Kociumaka, Ritu Kundu, Jakub Radoszewski, Wojciech Rytter, and Tomasz Waleń. How much different are two words with different shortest periods. In Lazaros S. Iliadis, Ilias Maglogiannis, and Vassilis P. Plagianakos, editors, *14th International Conference on Artificial Intelligence Applications and Innovations, AIAI 2018, Workshops*, volume 520 of *IFIP Advances in Information and Communication Technology*, pages 168–178. Springer, 2018. doi:10.1007/978-3-319-92016-0\_16.
- 4 Amihod 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.
- 5 Alexandr Andoni, Assaf Goldberger, Andrew McGregor, and Ely Porat. Homomorphic fingerprints under misalignments: sketching edit and shift distances. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *45th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2013*, pages 931–940. ACM, 2013. doi:10.1145/2488608.2488726.
- 6 Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In Shang-Hua Teng, editor, *19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, pages 343–352. SIAM, 2008. URL: <http://dl.acm.org/citation.cfm?id=1347082.1347120>.
- 7 Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Polylogarithmic approximation for edit distance and the asymmetric query complexity. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010*, pages 377–386. IEEE Computer Society, 2010. doi:10.1109/FOCS.2010.43.
- 8 Ziv Bar-Yossef, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. The sketching complexity of pattern matching. In Klaus Jansen, Sanjeev Khanna, José D. P. Rolim, and Dana Ron, editors, *8th International Workshop on Randomization and Computation, RANDOM 2004*, volume 3122 of *LNCS*, pages 261–272. Springer, 2004. doi:10.1007/978-3-540-27821-4\_24.
- 9 Djamal Belazzougui and Qin Zhang. Edit distance: Sketching, streaming, and document exchange. In Irit Dinur, editor, *57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*, pages 51–60. IEEE Computer Society, 2016. doi:10.1109/FOCS.2016.15.
- 10 Karl Bringmann, Marvin Künnemann, and Philip Wellnitz. Few matches or almost periodicity: Faster pattern matching with mismatches in compressed texts. In Timothy M. Chan, editor, *30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 1126–1145. SIAM, 2019. doi:10.1137/1.9781611975482.69.
- 11 Diptarka Chakraborty, Elazar Goldenberg, and Michal Koucký. Streaming algorithms for embedding and computing edit distance in the low distance regime. In Daniel Wichs and Yishay Mansour, editors, *48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016*, pages 712–725. ACM, 2016. doi:10.1145/2897518.2897577.
- 12 Timothy M. Chan, Shay Golan, Tomasz Kociumaka, Tsvi Kopelowitz, and Ely Porat. Approximating text-to-pattern Hamming distances. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020*, pages 643–656. ACM, 2020. doi:10.1145/3357713.3384266.
- 13 Panagiotis Charalampopoulos, Tomasz Kociumaka, and Philip Wellnitz. Faster approximate pattern matching: A unified approach, 2020. arXiv:2004.08350.
- 14 Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana Starikovskaya. The  $k$ -mismatch problem revisited. In Robert Krauthgamer, editor, *27th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016*, pages 2039–2052. SIAM, 2016. doi:10.1137/1.9781611974331.ch142.

- 15 Raphaël Clifford, Tomasz Kociumaka, and Ely Porat. The streaming  $k$ -mismatch problem. In Timothy M. Chan, editor, *30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 1106–1125. SIAM, 2019. doi:10.1137/1.9781611975482.68.
- 16 Raphaël Clifford and Tatiana Starikovskaya. Approximate Hamming distance in a stream. 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 20:1–20:14. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPICs.ICALP.2016.20.
- 17 Graham Cormode. Data sketching. *Communications of the ACM*, 60(9):48–55, 2017. doi:10.1145/3080008.
- 18 Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1–3):1–294, 2011. doi:10.1561/19000000004.
- 19 Michael S. Crouch and Andrew McGregor. Periodicity and cyclic shifts via linear sketches. In Leslie Ann Goldberg, Klaus Jansen, R. Ravi, and José D. P. Rolim, editors, *14th International Workshop on Approximation Algorithms for Combinatorial Optimization, APPROX 2011*, volume 6845 of *LNCS*, pages 158–170. Springer, 2011. doi:10.1007/978-3-642-22935-0\_14.
- 20 Benjamin Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. In *Natural Computing Series*, pages 1–87. Springer International Publishing, 2020. doi:10.1007/978-3-030-29414-4\_1.
- 21 Michael J. Fischer and Michael S. Paterson. String matching and other products. In Richard M. Karp, editor, *Complexity of Computation*, volume 7 of *SIAM-AMS Proceedings*, pages 113–125. AMS, 1974.
- 22 Paweł Gawrychowski and Przemysław Uznański. Towards unified approximate pattern matching for Hamming and  $L_1$  distance. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, volume 107 of *LIPICs*, pages 62:1–62:13. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICs.ICALP.2018.62.
- 23 Shay Golan, Tsvi Kopelowitz, and Ely Porat. Towards optimal approximate streaming pattern matching by matching multiple patterns in multiple streams. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, volume 107 of *LIPICs*, pages 65:1–65:16. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICs.ICALP.2018.65.
- 24 Shay Golan, Tsvi Kopelowitz, and Ely Porat. Streaming pattern matching with  $d$  wildcards. *Algorithmica*, 81(5):1988–2015, 2019. doi:10.1007/s00453-018-0521-7.
- 25 Richard W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950. doi:10.1002/j.1538-7305.1950.tb00463.x.
- 26 Wei Huang, Yaoyun Shi, Shengyu Zhang, and Yufan Zhu. The communication complexity of the Hamming distance problem. *Information Processing Letters*, 99(4):149–153, 2006. doi:10.1016/j.ipl.2006.01.014.
- 27 Howard J. Karloff. Fast algorithms for approximately counting mismatches. *Information Processing Letters*, 48(2):53–60, 1993. doi:10.1016/0020-0190(93)90177-B.
- 28 Richard M. Karp and Michael O. Rabin. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2):249–260, 1987. doi:10.1147/rd.312.0249.
- 29 Subhash Khot and Assaf Naor. Nonembeddability theorems via fourier analysis. *Mathematische Annalen*, 334(4):821–852, 2006. doi:10.1007/s00208-005-0745-0.
- 30 Tsvi Kopelowitz and Ely Porat. Breaking the variance: Approximating the Hamming distance in  $1/\varepsilon$  time per alignment. In Venkatesan Guruswami, editor, *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, pages 601–613. IEEE Computer Society, 2015. doi:10.1109/FOCS.2015.43.

- 31 Tsvi Kopelowitz and Ely Porat. A simple algorithm for approximating the text-to-pattern Hamming distance. In Raimund Seidel, editor, *1st Symposium on Simplicity in Algorithms, SOSA 2018*, volume 61 of *OASICS*, pages 10:1–10:5. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi:10.4230/OASICS.SOSA.2018.10.
- 32 S. Rao Kosaraju. Efficient string matching. Manuscript, 1987.
- 33 Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000. doi:10.1137/S0097539798347177.
- 34 Jelani Nelson. Sketching and streaming algorithms for processing massive data. *ACM Crossroads*, 19(1):14–19, 2012. doi:10.1145/2331042.2331049.
- 35 Rafail Ostrovsky and Yuval Rabani. Low distortion embeddings for edit distance. *Journal of the ACM*, 54(5):23, 2007. doi:10.1145/1284320.1284322.
- 36 Benny Porat and Ely Porat. Exact and approximate pattern matching in the streaming model. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009*, pages 315–323. IEEE Computer Society, 2009. doi:10.1109/FOCS.2009.11.
- 37 Ely Porat and Ohad Lipsky. Improved sketching of Hamming distance with error correcting. In Bin Ma and Kaizhong Zhang, editors, *18th Annual Symposium on Combinatorial Pattern Matching, CPM 2007*, volume 4580 of *LNCS*, pages 173–182. Springer, 2007. doi:10.1007/978-3-540-73437-6\_19.
- 38 Jakub Radoszewski and Tatiana Starikovskaya. Streaming  $k$ -mismatch with error correcting and applications. *Information and Computation*, 271:104513, 2020. doi:10.1016/j.ic.2019.104513.
- 39 Tatiana Starikovskaya, Michal Svagerka, and Przemysław Uznański.  $L_p$  pattern matching in a stream. In Jarosław Byrka and Raghu Meka, editors, *23rd International Workshop on Approximation Algorithms for Combinatorial Optimization, APPROX 2020*, volume 176 of *LIPICs*, pages 35:1–35:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.APPROX/RANDOM.2020.35.
- 40 David P. Woodruff. Optimal space lower bounds for all frequency moments. In J. Ian Munro, editor, *15th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004*, pages 167–175. SIAM, 2004. URL: <http://dl.acm.org/citation.cfm?id=982792.982817>.

## A Efficient Shift Distance Decoders

In this section, we develop exact and approximate  $k$ -mismatch shift distance sketches with efficient decoding procedures. These sketches use the same encoding functions as the corresponding  $k$ -mismatch circular sketches, so we only need to develop the decoding procedures.

Our decoding procedures for shift distance heavily rely on their counterparts for decoding the Hamming distance between  $S_1$  and a fixed rotation of  $S_2$ . Hence, each of the following four propositions refers to its counterpart in Section 4 or Section 6.

### A.1 Shift Distance Sketches for Non-Pseudo-Periodic Strings

► **Proposition A.1** (see Proposition 4.5). *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon,k}$  of Construction 4.4, forms an  $(\varepsilon, k)$ -ASDS sketch of  $\Sigma^n \setminus \mathcal{H}_{n,k}$ . The decoding algorithm costs  $\tilde{O}(\varepsilon^{-2}k)$  time with high probability.*

**Proof.** Our decoding procedure iterates over  $i \in f(S_1) \cap A$  and  $i' \in f(S_2) \cap B$ . For each such pair  $(i, i')$ , the procedure retrieves the sketches  $\text{sk}_{\varepsilon}(\text{cyc}^i(S_1))$  and  $\text{sk}_{\varepsilon}(\text{cyc}^{i'}(S_2))$  and recovers a  $(1 + \varepsilon)$ -approximation of  $\text{Ham}(\text{cyc}^i(S_1), \text{cyc}^{i'}(S_2))$ . The algorithm returns the smallest among the values obtained across all the iterations.



Since  $\text{Ham}(\text{cyc}^i(S_1), \text{cyc}^{i'}(S_2)) \geq \text{sh}(S_1, S_2)$ , the returned value is at least  $(1 - \varepsilon)\text{sh}(S_1, S_2)$  with high probability (unless the sketches  $\text{sk}_\varepsilon$  fail). Moreover, if  $\text{sh}(S_1, S_2) \leq k$  with  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \text{sh}(S_1, S_2)$  for some integer  $m$ , then, as in the proof of Proposition 4.5, with high probability, there is a pair of indices  $i \in f(S_1) \cap A$  and  $i' \in f(S_2) \cap B$  with  $i' = (i + m) \circlearrowleft n$ . Hence, the returned value is at most  $(1 + \varepsilon)\text{Ham}(S_1, \text{cyc}^m(S_2)) = (1 + \varepsilon)\text{sh}(S_1, S_2)$  with high probability.  $\blacktriangleleft$

► **Proposition A.2** (see Proposition 4.8). *There exists a decoding function which, together with the encoding  $\text{circ}_k$  of Construction 4.7, forms a  $k$ -ESDS sketch of  $\Sigma^n \setminus \mathcal{H}_{n,k}$ . The decoding algorithm costs  $\tilde{O}(k^2)$  time with high probability.*

**Proof.** The decoding algorithm first computes the sizes  $s_m := |f(S_1) \cap \text{rot}_n^m(f(S_2))|$  for all shifts  $m \in [n]$ . For this, the algorithm iterates over  $i \in f(S_1)$  and  $i' \in f(S_2)$  incrementing  $s_{(i'-i) \circlearrowleft n}$ . Next, for each shift  $m \in [n]$  with  $c_m \geq k$ , the algorithm uses the decoding function of Proposition 4.8 to retrieve  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  (or learn that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) > k$ ). Finally, the algorithm returns the smallest among the reported values. (If  $c_m < k$  for each  $m \in [n]$ , then the algorithm returns  $\infty$ .)

As for correctness, first note that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \geq \text{sh}(S_1, S_2)$  holds for each  $m \in [n]$ , so the returned value is at least  $\min(k + 1, \text{sh}(S_1, S_2))$  with high probability (unless the decoding procedure of Proposition 4.8 fails). Next, suppose that  $\text{sh}(S_1, S_2) = \text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$ . As argued in the proof of Proposition 4.8,  $s_m = |f(S_1) \cap f(\text{cyc}^m(S_1))| \geq k$  holds with high probability. Consequently, the decoding procedure of Proposition 4.8 was called for  $S_1, S_2$ , and  $m$ , resulting in  $\text{sh}(S_1, S_2)$  with high probability. Hence, the returned value is at most  $\text{sh}(S_1, S_2)$  with high probability.

The decoder iterates over  $f(S_1) \times f(S_2)$ , which is of size  $\tilde{O}(k^2)$  with high probability due to Theorem 4.2. Hence, by the pigeonhole principle there are at most  $\tilde{O}(\frac{k^2}{k}) = \tilde{O}(k)$  positions  $m \in [n]$  such that  $c_m \geq k$ . For each such position, the decoding time of Proposition 4.8 is  $\tilde{O}(k)$ . Thus, the total decoding time is  $\tilde{O}(k^2)$ .  $\blacktriangleleft$

## A.2 Shift Distance Sketches for Pseudo-Periodic Strings

► **Lemma A.3** (see Lemma 6.5). *There exists a decoding function  $\text{dec}_0^{\text{sh}}$  which, together with the encoding  $\text{sk}_0$  of Lemma 6.5, forms an exact 0-ESDS sketch with constant decoding time.*

**Proof.** The decoding function, given the sketches  $\text{sk}_0(S_1) = (\Phi(\text{minrot}(S_1)), \text{root}(S_1), r_1)$  and  $\text{sk}_0(S_2) = (\Phi(\text{minrot}(S_2)), \text{root}(S_2), r_2)$ , returns 0 or  $\infty$  based on whether  $\Phi(\text{minrot}(S_1)) = \Phi(\text{minrot}(S_2))$  or not.  $\blacktriangleleft$

► **Proposition A.4** (see Proposition 6.7). *There exists a decoding function which, together with the encoding  $\text{circ}_k$  of Construction 6.6, forms a  $k$ -ESDS sketch of  $\mathcal{H}'_{n,k}$ . The decoding algorithm costs  $\tilde{O}(k^2)$  time with high probability.*

**Proof.** The decoding algorithm is given the sketches  $\text{circ}_k(S_1) = (\text{sk}_0(S'_1), \text{MI}(S_1, S'_1))$  and  $\text{circ}_k(S_2) = (\text{sk}_0(S'_2), \text{MI}(S_2, S'_2))$ . First, the algorithm applies  $\text{dec}_0^{\text{sh}}(\text{sk}_0(S'_1), \text{sk}_0(S'_2))$  of Lemma A.3. If this call returns a non-zero result, then  $\infty$  is returned. Otherwise, for each  $m \in [n]$ , the algorithm constructs the following sets:

$$\begin{aligned} P_m &:= \text{MP}(S_1, S'_1) \cap \text{MP}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2)) \\ P'_m &:= P_m \setminus \text{MP}(S_1, \text{cyc}^m(S_2)) \end{aligned}$$

For this, the algorithm iterates over  $(i, a, b) \in \text{MI}(S_1, S'_1)$  and  $(i', c, d) \in \text{MP}(S_2, S'_2)$ , adding  $i$  to  $P_{(i'-i) \circlearrowleft n}$  and, provided that  $a = c$ , also to  $P'_{(i'-i) \circlearrowleft n}$ .



For each shift  $m$  with  $P_m \neq \emptyset$ , the algorithm uses  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$  of Lemma 6.5. If this call returns a non-zero result, then  $m$  is ignored. Otherwise,  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) - |P_m| - |P'_m|$  is computed. Finally, the algorithm returns the minimum of  $\text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2)$  and the smallest among the computed values  $\text{Ham}(S_1, \text{cyc}^m(S_2))$ .

**Correctness.** By Corollary 6.3,  $\text{sh}(S_1, S_2) \leq k$  guarantees  $\text{sh}(S'_1, S'_2) = 0$ , so the algorithm correctly returns  $\infty$  if  $\text{dec}_0^{\text{sh}}(\text{sk}_0(S'_1), \text{sk}_0(S'_2))$  yields a non-zero result. Moreover,  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$  guarantees  $S'_1 = \text{cyc}^m(S'_2)$ , so the algorithm correctly ignores  $m \in [n]$  if  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$  yields a non-zero result. In the following, we assume  $\text{sh}(S'_1, S'_2) = 0$  with  $S'_1 = \text{cyc}^m(S'_2)$  for all the shifts considered. The latter assumption implies  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) - |P_m| - |P'_m|$  (compare the proof of Fact 3.2). Moreover,  $\text{sh}(S_1, S_2) \leq \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2)$  holds by the triangle inequality. Hence, the returned value is at least  $\text{sh}(S_1, S_2)$  with high probability.

On the other hand, if  $\text{sh}(S_1, S_2) = \text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$  for some shift  $m \in [n]$ , then  $S'_1 = \text{cyc}^m(S'_2)$  and  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) - |P_m| - |P'_m|$ . This either yields  $\text{sh}(S_1, S_2) = \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2)$  (in case of  $P_m = \emptyset$ , which yields  $|P_m| = |P'_m| = 0$ ) or that  $m$  was among the shifts considered (otherwise). In both cases, we conclude that the returned value is at most  $\text{sh}(S_1, S_2)$  with high probability. ◀

A relaxed  $(\varepsilon, k)$ -ASDS sketch is defined analogously to a relaxed  $(\varepsilon, k)$ -ACS sketch:

- if  $\text{sh}(S_1, S_2) < \frac{1}{2}k$ , then  $\text{dec}^{\text{sh}}(\text{sk}(S_1), \text{sk}(S_2)) < \frac{1+\varepsilon}{2}k$ ,
- if  $\frac{1}{2}k \leq \text{sh}(S_1, S_2) \leq k$ , then  $\text{dec}^{\text{sh}}(\text{sk}(S_1), \text{sk}(S_2)) \in (1 \pm \varepsilon)\text{sh}(S_1, S_2)$ ,
- otherwise,  $\text{dec}^{\text{sh}}(\text{sk}(S_1), \text{sk}(S_2)) > (1 - \varepsilon)k$ .

► **Proposition A.5** (see Proposition 6.9). *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon, k}$  of Construction 6.8, forms a relaxed  $(\varepsilon, k)$ -ASDS sketch of  $\mathcal{H}'_{n, k}$ . The decoding algorithm costs  $\tilde{O}(\varepsilon^{-2}k)$  time with high probability.*

**Proof.** The decoding algorithm is given sketches  $\text{circ}_{\varepsilon, k}(S_1) = (\text{sk}_0(S'_1), \text{MI}_{A \cup B}(S_1, S'_1))$  and  $\text{circ}_{\varepsilon, k}(S_2) = (\text{sk}_0(S'_2), \text{MI}_{A \cup B}(S_2, S'_2))$ . First, the algorithm applies  $\text{dec}_0^{\text{sh}}(\text{sk}_0(S'_1), \text{sk}_0(S'_2))$  of Lemma A.3. If this call returns a non-zero result, then  $\infty$  is returned. Otherwise, for each  $m \in [n]$ , the algorithm constructs the sets  $P_m \cap A \cap \text{rot}_n^m(B)$  and  $P'_m \cap A \cap \text{rot}_n^m(B)$ , where

$$P_m := \text{MP}(S_1, S'_1) \cap \text{MP}(\text{cyc}^m(S_2), \text{cyc}^m(S'_2))$$

$$P'_m := P_m \setminus \text{MP}(S_1, \text{cyc}^m(S_2))$$

are defined as in the proof of Proposition A.4. For this, the algorithm iterates over  $(i, a, b) \in \text{MI}_A(S_1, S'_1)$  and  $(i', c, d) \in \text{MI}_B(S_2, S'_2)$ , adding  $i$  to  $P_{(i'-i) \circledast n} \cap A \cap \text{rot}_n^{(i'-i) \circledast n}(B)$  and, provided that  $a = c$ , also to  $P'_{(i'-i) \circledast n} \cap A \cap \text{rot}_n^{(i'-i) \circledast n}(B)$ .

For each shift  $m$  with  $P_m \cap A \cap \text{rot}_n^m(B) \neq \emptyset$ , the algorithm uses  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$  of Lemma 6.5. If this call returns a non-zero result, then  $m$  is ignored. Otherwise, the algorithm computes

$$d_m := \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) - \frac{\varepsilon^2 k}{\log n} (|P_m \cap A \cap \text{rot}_n^m(B)| + |P'_m \cap A \cap \text{rot}_n^m(B)|).$$

Finally, the algorithm returns the minimum of  $\text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2)$  and the smallest among the computed values  $d_m$ .

**Correctness.** By Corollary 6.3,  $\text{sh}(S_1, S_2) \leq k$  guarantees  $\text{sh}(S'_1, S'_2) = 0$ , so the algorithm correctly returns  $\infty$  if  $\text{dec}_0^{\text{sh}}(\text{sk}_0(S'_1), \text{sk}_0(S'_2))$  yields a non-zero result. Moreover,  $\text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$  guarantees  $S'_1 = \text{cyc}^m(S'_2)$ , so the algorithm correctly ignores  $m \in [n]$  if  $\text{dec}_0(\text{sk}_0(S'_1), \text{sk}_0(S'_2), m)$  yields a non-zero result. In the following, we assume  $\text{sh}(S'_1, S'_2) = 0$  with  $S'_1 = \text{cyc}^m(S'_2)$  for all the shifts considered.

Recall that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) - |P_m| - |P'_m|$  holds provided that  $S'_1 = \text{cyc}^m(S'_2)$ . Since  $A \cap \text{rot}_n^m(B)$  is a random subset of  $[n]$  with elements sampled independently with rate  $\frac{\log n}{\varepsilon^2 k}$ , the quantity  $\frac{\varepsilon^2 k}{\log n} (|P_m \cap A \cap \text{rot}_n^m(B)| + |P'_m \cap A \cap \text{rot}_n^m(B)|)$  is with high probability a  $\pm \frac{\varepsilon k}{2}$ -additive approximation of  $|P_m| + |P'_m|$  (which can be argued as in the proof of Lemma 3.1). Consequently, the computed value  $d_m$  is with high probability a  $\pm \frac{\varepsilon k}{2}$ -additive approximation of  $\text{Ham}(S_1, \text{cyc}^m(S_2))$ . As  $\text{sh}(S_1, S_2) \leq \text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2)$  holds by the triangle inequality, this means that the returned value is at least  $(1 - \varepsilon)\text{sh}(S_1, S_2)$  with high probability provided that  $\text{sh}(S_1, S_2) \geq \frac{1}{2}k$ .

On the other hand, if  $\text{sh}(S_1, S_2) = \text{Ham}(S_1, \text{cyc}^m(S_2)) \leq k$  for some shift  $m \in [n]$ , then  $S'_1 = \text{cyc}^m(S'_2)$  and  $d_m$  is a  $\pm \frac{\varepsilon k}{2}$ -additive approximation of  $\text{sh}(S_2, S_2)$ . This either yields  $\text{Ham}(S_1, S'_1) + \text{Ham}(S_2, S'_2) \leq (1 + \varepsilon)\text{sh}(S_1, S_2)$  (if  $P_m \cap A \cap \text{rot}_n^m(B) = \emptyset$ , which yields  $|P_m \cap A \cap \text{rot}_n^m(B)| = |P'_m \cap A \cap \text{rot}_n^m(B)| = 0$ ) or that  $m$  was among the shifts considered (otherwise). In both cases, we conclude that the returned value is at most  $(1 + \varepsilon)\text{sh}(S_1, S_2)$  with high probability.  $\blacktriangleleft$

### A.3 Shift Distance Sketches for $\Sigma^n$

After handling non-pseudo-periodic and pseudo-periodic strings separately, we derive sketches for the whole  $\Sigma^n$ . The following results provide efficient shift distance decoding procedures for the circular  $k$ -mismatch sketches described in Section 7.

For the exact case, using Propositions A.2 and A.4, the same construction as in the proof of Theorem 1.3 yields the following corollary.

► **Corollary A.6** (see Theorem 1.3). *There exists a  $k$ -ESDS sketch of size  $\tilde{O}(k)$  with decoding time  $\tilde{O}(k^2)$ .*

For the approximate case, using Propositions A.1 and A.5, the same construction as in Proposition 7.1 yields a relaxed  $(\varepsilon, k)$ -ASDS sketch of size  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$  with decoding time  $\tilde{O}(\varepsilon^{-2}k)$ .

► **Proposition A.7** (see Proposition 7.1). *There exists a relaxed  $(\varepsilon, k)$ -ASDS sketch of size  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$  and decoding time of  $\tilde{O}(\varepsilon^{-2}k)$ .*

The following provides an alternative method for constructing  $(\varepsilon, k)$ -ASDS sketches which improves the sketch size (for  $k \geq \varepsilon n$ ) but degrades the decoding time.

► **Proposition A.8** (see Proposition 7.3). *There exists a decoding function which, together with the encoding  $\text{circ}_{\varepsilon, k}$  of Construction 7.2, forms a relaxed  $(\varepsilon, k)$ -ASDS sketch of  $\Sigma^n$ . The decoding time is  $\tilde{O}(\frac{n^2}{\varepsilon^2 k})$  with high probability.*

**Proof.** The decoding function, given the sketches of  $S_1, S_2 \in \Sigma^n$ , computes a value  $d_m := \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  for each  $m \in [n]$ . For this, the algorithm iterates over  $(i, S_1[i])$  with  $i \in A$  (retrieved from the sketch of  $S_1$ ) and  $(i', S_2[i'])$  with  $i' \in B$  (retrieved from the sketch of  $S_2$ ), and increments  $d_{(i' - i) \circ n}$  if  $S_1[i] \neq S_2[i']$ .

As in the proof of Proposition 7.3,  $\frac{\varepsilon^2 k}{\log n} \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  is a  $(1 \pm \varepsilon)$ -approximation of  $\text{Ham}(S_1, \text{cyc}^m(S_2))$  with high probability provided that  $\text{Ham}(S_1, \text{cyc}^m(S_2)) = \Omega(k)$  (and  $\frac{\varepsilon^2 k}{\log n} \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2)) = o(k)$  otherwise). Hence, the algorithm returns as an approximation of  $\text{sh}(S_1, S_2)$  the smallest value  $\frac{\varepsilon^2 k}{\log n} \text{Ham}_{A \cap \text{rot}_n^m(B)}(S_1, \text{cyc}^m(S_2))$  among  $m \in [n]$ .  $\blacktriangleleft$

► **Corollary A.9** (see Theorem 1.4). *There exists an  $(\varepsilon, k)$ -ASDS sketch of size  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$  with decoding time  $\tilde{O}(\varepsilon^{-2}k)$ , and an  $(\varepsilon, k)$ -ASDS sketch of size  $\tilde{O}(\varepsilon^{-1.5}\sqrt{n})$  with decoding time  $\tilde{O}(\varepsilon^{-3}n)$ .*

## B Missing Proofs

► **Lemma 3.1.** *Let  $A$  be a random subset of  $[n]$  with elements chosen independently at rate  $p$ . For  $0 < \varepsilon < 1$ , we have  $\Pr[\text{Ham}_A(S, T) \in (1 \pm \varepsilon)p\text{Ham}(S, T)] \geq 1 - 2 \exp\left(-\frac{p\text{Ham}(S, T)\varepsilon^2}{3}\right)$ .*

**Proof.** For each index  $i \in [n]$ , let  $x_i$  be an indicator variable such that  $x_i = 1$  if  $i \in \text{MI}_A(S, T)$  and  $x_i = 0$  otherwise. Note that  $\text{Ham}_A(S, T) = |\text{MI}_A(S, T)| = \sum_{i=1}^n x_i$  and that  $x_i$  are independent variables. For every  $i \in \text{MI}(S, T)$ , we have  $\Pr[x_i = 1] = p$  and, for every  $i \notin \text{MI}(S, T)$ , we have  $\Pr[x_i = 1] = 0$ . Thus,  $\mathbb{E}[\sum_{i=1}^n x_i] = \mathbb{E}[\sum_{i \in \text{MI}_A(S, T)} x_i] = p\text{Ham}_A(S, T)$ . Hence, by the Chernoff bound (see, e.g., [20])

$$\Pr\left[\left|\sum_{i=1}^n x_i - p\text{Ham}_A(S, T)\right| > \varepsilon p\text{Ham}_A(S, T)\right] \leq 2 \exp\left(-\frac{p\text{Ham}_A(S, T)\varepsilon^2}{3}\right).$$

Thus,  $\Pr[\text{Ham}_A(S, T) \in (1 \pm \varepsilon)p\text{Ham}(S, T)] \geq 1 - 2 \exp\left(-\frac{p\text{Ham}(S, T)\varepsilon^2}{3}\right)$ .  $\blacktriangleleft$

► **Fact 3.2.** *For every  $S, T, U \in \Sigma^n$  and every  $A \subseteq [n]$ , the mismatch information  $\text{MI}_A(S, U)$  can be retrieved from  $\text{MI}_A(S, T)$  and  $\text{MI}_A(T, U)$  in time  $\tilde{O}(\text{Ham}_A(S, T) + \text{Ham}_A(T, U))$ .*

**Proof.** For each  $i \in A$ , we have one of the following four cases:

- if  $i \notin \text{MP}(S, T)$  and  $i \notin \text{MP}(T, U)$ , then  $S[i] = T[i] = U[i]$ , so  $i \notin \text{MP}(S, U)$ ,
- if  $(i, a, b) \in \text{MI}(S, T)$  and  $i \notin \text{MP}(T, U)$ , then  $S[i] = a \neq b = T[i] = U[i]$ , so  $(i, a, b) \in \text{MI}(S, U)$ ,
- if  $i \notin \text{MP}(S, T)$  and  $(i, b, c) \in \text{MP}(T, U)$ , then  $S[i] = T[i] = b \neq c = U[i]$ , so  $(i, b, c) \in \text{MI}(S, U)$ ,
- if  $(i, a, b) \in \text{MI}(S, T)$  and  $(i, b, c) \in \text{MP}(T, U)$ , then  $S[i] = a \neq b = T[i] = b \neq c = U[i]$ , so  $(i, a, c) \in \text{MI}(S, U)$  (if  $a \neq c$ ) or  $i \notin \text{MP}(S, U)$  (if  $a = c$ ).  $\blacktriangleleft$

► **Theorem 4.3** ( $(1 \pm \varepsilon)$ -approximate sketches, folklore). *There exists a  $(1 \pm \varepsilon)$ -approximate sketch  $\text{sk}_\varepsilon$  such that, given  $\text{sk}_\varepsilon(S_1)$  and  $\text{sk}_\varepsilon(S_2)$  for two strings  $S_1, S_2 \in \Sigma^n$ , one can decode  $\text{Ham}(S_1, S_2)$  with a  $(1 \pm \varepsilon)$ -multiplicative error. The sketches use  $\tilde{O}(\varepsilon^{-2})$  space, the decoding algorithm is correct with high probability and costs  $\tilde{O}(\varepsilon^{-2})$  time.*

**Proof.** Consider  $\mu : \Sigma \rightarrow \{0, 1\}^\sigma$  defined as  $\mu(c) = 0^{c-1}10^{\sigma-c}$ . For every words  $u, v$ , we have  $\text{Ham}(\mu(u), \mu(v)) = 2 \cdot \text{Ham}(u, v)$ . We then use AMS sketches [2] on  $\mu(u)$  and  $\mu(v)$  which allow for decoding of  $\ell_2$  distance  $\|\mu(u) - \mu(v)\|_2$ . This is enough since, for binary words, the  $\ell_2^2$  distance coincides with the Hamming distance. We then note that the AMS sketches of  $\mu(u)$  and  $\mu(v)$  can be computed without explicitly constructing  $\mu(u)$  or  $\mu(v)$ .  $\blacktriangleleft$

► **Proposition 7.1.** *There exists a relaxed  $(\varepsilon, k)$ -ACS sketch for  $\Sigma^n$  of size  $\tilde{O}(\varepsilon^{-2}\sqrt{k})$ . Its decoding time is  $\tilde{O}(\varepsilon^{-1}\sqrt{k} + \varepsilon^{-2})$  with high probability.*

## 46:24 Improved Circular $k$ -Mismatch Sketches

**Proof.** Our construction combines the  $(\varepsilon, k)$ -ACS sketch of Proposition 4.5 and the relaxed  $(\varepsilon, k)$ -ACS sketch of Proposition 6.9. For each strings  $S$ , if  $S \in \mathcal{H}'_{n,k}$ , then the sketch contains the sketch of  $S$  by Proposition 6.9 and, if  $S \in \Sigma^n \setminus \mathcal{H}_{n,k}$ , then the sketch contains the sketch of  $S$  by Proposition 4.5. Notice that, for  $S \in \mathcal{H}'_{n,k} \setminus \mathcal{H}_{n,k}$  the sketch contains both components.

For any two strings  $S_1, S_2 \in \Sigma^n$ , given the sketches of  $S_1$  and  $S_2$ , the decoder works as follows. If the two sketches contains compatible components (of Proposition 4.5 or of Proposition 6.9), then the decoder uses the decoder corresponding to these components. Otherwise, without loss of generality, it must be that  $S_1 \in \mathcal{H}_{n,k}$  and  $S_2 \notin \mathcal{H}'_{n,k}$ . Thus, by Observation 6.2,  $\text{Ham}(S_1, S_2) > k$ , and therefore the decoder outputs  $\infty$ .  $\blacktriangleleft$