


Matching Patterns with Variables Under Hamming Distance

Paweł Gawrychowski ✉ 

Faculty of Mathematics and Computer Science, University of Wrocław, Poland

Florin Manea ✉ 

Computer Science Department and Campus-Institut Data Science, Göttingen University, Germany

Stefan Siemer ✉ 

Computer Science Department, Göttingen University, Germany

Abstract

A pattern α is a string of variables and terminal letters. We say that α matches a word w , consisting only of terminal letters, if w can be obtained by replacing the variables of α by terminal words. The matching problem, i.e., deciding whether a given pattern matches a given word, was heavily investigated: it is NP-complete in general, but can be solved efficiently for classes of patterns with restricted structure. In this paper, we approach this problem in a generalized setting, by considering approximate pattern matching under Hamming distance. More precisely, we are interested in what is the minimum Hamming distance between w and any word u obtained by replacing the variables of α by terminal words. Firstly, we address the class of regular patterns (in which no variable occurs twice) and propose efficient algorithms for this problem, as well as matching conditional lower bounds. We show that the problem can still be solved efficiently if we allow repeated variables, but restrict the way the different variables can be interleaved according to a locality parameter. However, as soon as we allow a variable to occur more than once and its occurrences can be interleaved arbitrarily with those of other variables, even if none of them occurs more than once, the problem becomes intractable.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Theory of computation → Formal languages and automata theory

Keywords and phrases Pattern with variables, Matching algorithms, Hamming distance, Conditional lower bounds, Patterns with structural restrictions

Digital Object Identifier 10.4230/LIPIcs.MFCS.2021.48

Related Version *Full Version*: <https://arxiv.org/abs/2106.06249> [30]

Funding The work of the two authors from Göttingen was supported by the DFG-grant 389613931.

1 Introduction

A *pattern* (with variables) is a string which consists of *terminal letters* (e.g., a, b, c), treated as constants, and *variables* (e.g., x_1, x_2). A pattern is mapped to a word by substituting the variables by strings of terminals. For example, $x_1x_1babx_2x_2$ can be mapped to **aaaababbb** by the substitution ($x_1 \rightarrow \mathbf{aa}, x_2 \rightarrow \mathbf{b}$). If a pattern α can be mapped to a string of terminals w , we say that α matches w . The problem of deciding whether there exists a substitution which maps a given pattern α to a given word w is called the *matching problem*.

Patterns with variables and their matching problem appear in various areas of theoretical computer science. In particular, the matching problem is a particular case of the satisfiability problem for word equations. These are equations whose both sides are patterns with variables and whose solutions are substitutions that map both sides to the same word [37]; in the pattern matching problem, one side of the input equation is a string of terminals. Patterns with variables occur also in combinatorics on words (e.g., unavoidable patterns [38]), stringology (e.g., generalized function matching [2]), language theory (e.g., pattern languages [3]), or



© Paweł Gawrychowski, Florin Manea, and Stefan Siemer;
licensed under Creative Commons License CC-BY 4.0

46th International Symposium on Mathematical Foundations of Computer Science (MFCS 2021).

Editors: Filippo Bonchi and Simon J. Puglisi; Article No. 48; pp. 48:1–48:24

Leibniz International Proceedings in Informatics



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

database theory (e.g., document spanners [27, 26, 19, 44]). In a more practical setting, patterns with variables are used in connection to extended regular expressions with backreferences [14, 29, 25, 28], used in various programming languages.

The *matching problem* is NP-complete [3] in general. This is especially unfortunate for some computational tasks on patterns which implicitly solve the matching problem and are thus intractable as well. For instance, in algorithmic learning theory, this is the case for the task of computing *descriptive patterns* for finite sets of words [3, 21]. Such descriptive patterns are useful for the inductive inference of pattern languages, a prominent example of a language class which can be inferred from positive data (see, the survey [46] and the references therein). This and many other applications of pattern matching provide a good motivation to identify cases in which the matching problem becomes tractable. A natural approach to this task is to consider restricted classes of patterns. A thorough analysis [42, 45, 23, 24, 22, 43] of the complexity of the matching problem has provided several subclasses of patterns for which the matching problem is in P, when some structural parameters of patterns are bounded by constants. Prominent examples in this direction are patterns with a bounded number of repeated variables occurring in a pattern, patterns with bounded scope coincidence degree [42], or patterns with bounded locality [18]. The formal definitions of these parameters are given in Section 4, and corresponding efficient matching algorithms be found in [22, 18], but, to give an intuition, we mention that they are all numerical parameters which describe the structure of patterns and parameterize the complexity of the matching algorithms. That is, in all cases, if the respective parameter equals k , the matching algorithm runs in $O(n^{ck})$ for some constant c , and, moreover, the matching problem can be shown to be $W[1]$ -hard w.r.t. the respective parameter. A more general approach [42] introduces the notion of treewidth of patterns, and shows that the matching problem can be solved in $O(n^{2k+4})$ time for patterns with bounded treewidth k . The algorithms resulting from this general theory are less efficient than the specialized ones, while the matching problem remains $W[1]$ -hard w.r.t. treewidth of patterns. See also the survey [39].

In this paper, we extend the study of patterns which can be matched efficiently to the case of approximate matching: we allow mismatches between the word w and the image of α under a substitution. More precisely, we consider two problems. In the decision problem MisMatch_P we are interested in deciding, for a given pattern α from a class P , a given word w , and a non-negative integer Δ whether there exists a variable-substitution h such that the word $h(\alpha)$ has at most Δ mismatches to the word w ; in other words, the Hamming distance $d_{\text{HAM}}(h(\alpha), w)$ between $h(\alpha)$ and w is at most Δ . Alternatively, we consider the corresponding minimisation problem MinMisMatch_P of computing $d_{\text{HAM}}(\alpha, w) = \min\{d_{\text{HAM}}(h(\alpha), w) \mid h \text{ is a substitution of the variables in } \alpha\}$.

As most real-world textual data (e.g., involving genetic data or text written by humans) contains errors, considering string-processing algorithms in an approximate setting is natural and has been heavily investigated. See, e.g., the recent papers [16, 32, 31, 47], and the references therein, as well as classical results such as [1, 41, 35]. Closer to the topic of this paper, the problem of approximate pattern matching was also considered in the context of regular expression matching – see [6, 41] and the references therein. Continuing this line of research, we initiate a study of approximate matching problems for patterns with variables. Intuitively, in our problems, we ask if the input word w is a few mismatches away from matching the pattern α , i.e., if w can be seen as a slightly erroneous version of a word which exactly matches α .

■ **Table 1** Our results are listed in columns 3 and 4. We assume $|w| = n$, $|\alpha| = m$, $|\text{var}(\alpha)| = p$.

Class	$\text{Match}(w, \alpha)$	$\text{MisMatch}(w, \alpha, \Delta)$	$\text{MinMisMatch}(w, \alpha)$
Reg	$O(n)$ [folklore]	$O(n\Delta)$ matching cond. lower bound	$O(nd_{\text{HAM}}(\alpha, w))$ matching cond. lower bound
1Var	$O(n)$ [folklore]	$O(n)$	$O(n)$
NonCross	$O(nm \log n)$ [22]	$O(n^3 p)$	$O(n^3 p)$
1RepVar $k = \# x\text{-blocks}$	$O(n^2)$ [22]	$O(n^{k+2} m)$ W[1]-hard w.r.t. k	$O(n^{k+2} m)$, PTAS W[1]-hard w.r.t. k no FPTAS (if $FPT \neq W[1]$)
kLOC	$O(mkn^{2k+1})$ [18] W[1]-hard w.r.t. k	$O(n^{2k+2} m)$ W[1]-hard w.r.t. k	$O(n^{2k+2} m)$ W[1]-hard w.r.t. k no FPTAS (if $FPT \neq W[1]$)
kSCD	$O(m^2 n^{2k})$ [22] W[1]-hard w.r.t. k	NP-hard for $k \geq 2$	NP-hard for $k \geq 2$
kRepVar	$O(n^{2k})$ [22] W[1]-hard w.r.t. k	NP-hard for $k \geq 1$	NP-hard for $k \geq 1$
k -bounded treewidth	$O(n^{2k+4})$ [42] W[1]-hard w.r.t. k	NP-hard for $k \geq 3$	NP-hard for $k \geq 3$

Our Contribution. Our results are summarized in Table 1. In that table, we describe the results we obtained for the problems MisMatch_P and MinMisMatch_P (introduced informally above and formally in Section 2) for a series of classes P of patterns for which the matching problem Match can be solved in polynomial time. The classes P we consider are the following: The class **Reg** of regular patterns, which contain at most one occurrence of any variable; the class **1Var** of unary patterns, which contain several occurrences of a single variable and terminals; the class **NonCross** of non-cross-patterns, which can be factorized in multiple **1Var**-patterns whose variables are pairwise different; the class **1RepVar** of one-repeated-variables, where only one variable (say x) is allowed to occur more than once; the classes **kLOC** of k -local patterns and **kSCD** of patterns with scope coincidence degree at most k , defined formally in Section 4; the class **kRepVar** of k -repeated-variables, where only k variables are allowed to occur more than once. We also (indirectly) obtain a lower bound for the complexity of MisMatch and MinMisMatch in the case of patterns with treewidth at most k .

Interestingly, for **Reg** we obtain matching upper and conditional lower bounds. As regular patterns are, in fact, a particular case of regular expressions, it is worth mentioning that, due to the conditional lower bounds from [4] on exact regular expression matching, it is not to be expected that the general case of matching regular-expressions under Hamming distance can be solved as efficiently as the case of regular patterns. Regarding patterns with repeated variables, we note that while in the case when the number of repeated variables, the scope coincidence degree, or the treewidth was bounded by a constant, polynomial-time algorithms for the exact matching problem were obtained. This does not hold in our approximate setting, unless $P=NP$. Only the locality measure has the same behaviour as in the case of exact matching: $\text{MisMatch}_{\text{kLOC}}$ and $\text{MinMisMatch}_{\text{kLOC}}$ can still be solved in polynomial time for constant k . In the simpler case of **1RepVar**-patterns, the locality corresponds to the number of x -blocks, so, if this is bounded by a constant, the two problems we consider can be solved in polynomial time.

The paper is organized as follows: after some preliminaries, we present in detail the results on **Reg**-patterns. Then we overview the results on patterns with repeated variables.

Future Work. While our results paint a detailed image of the complexity of `MisMatch` and `MinMisMatch` for some prominent classes of patterns for which the matching problem can be solved efficiently, some continuations of this work can be easily identified. Following [22], it would be interesting to try to optimise the algorithms for all classes from the table (except `Reg`, where the upper and conditional lower bounds match). In the case of `Reg`, it would be interesting to consider the problem for regular patterns with a constant number of variables; already in the case of two variables (also known as approximate string matching under Hamming distance) the known complexity upper and lower bounds do not match anymore [31, 47]. Another direction is to consider the two problems for other distance functions (e.g., edit distance) instead of the Hamming distance. Finally, it would be interesting if the applications of pattern matching in the area of algorithmic learning theory can be formulated (and still remain interesting) in this approximate setting.

2 Preliminaries

Let Σ be a finite alphabet of *terminal letters*. Let Σ^* be the set of all words and ε the empty word. The concatenation of k words w_1, w_2, \dots, w_k is written $\prod_{i=1}^k w_i$. The set Σ^+ is defined as $\Sigma^* \setminus \{\varepsilon\}$. For $w \in \Sigma^*$ the length of w is defined the number of symbols of w , and denoted as $|w|$. Further, let $\Sigma^n = \{w \in \Sigma^* \mid |w| = n\}$ and $\Sigma^{\leq n} = \bigcup_{i=0}^n \Sigma^i$. The letter on position i of w , for $1 \leq i \leq |w|$, is denoted by $w[i]$. For $w \in \Sigma^+$ and $x, y, z \in \Sigma^*$, the word y is a factor of w , if $w = xyz$; moreover, if $x = \varepsilon$ (respectively, $z = \varepsilon$), then y is called a prefix (respectively, suffix) of w . Let $w[i : j] = w[i] \cdots w[j]$ be the factor of w starting on position i and ending on position j ; if $i > j$ then $w[i : j] = \varepsilon$. By $[i : j]$ we denote the set $\{i, i + 1, \dots, j\}$ and $D[i : j]$ denotes a subarray of D whose positions are indexed by the numbers in $[i : j]$.

Let $\mathcal{X} = \{x_1, x_2, x_3, \dots\}$ be a set of *variables*. For the set of terminals Σ and the set of variables \mathcal{X} with $\Sigma \cap \mathcal{X} = \emptyset$, a pattern α is a word containing both terminals and variables, i.e., an element of $PAT_\Sigma = (\mathcal{X} \cup \Sigma)^+$. The set of all patterns, over all terminal-alphabets, is denoted $PAT = \bigcup_\Sigma PAT_\Sigma$. Given a word or a pattern γ , for the smallest sets (w.r.t. inclusion) $B \subseteq \Sigma$ and $Y \subseteq \mathcal{X}$ with $\gamma \in (B \cup Y)^*$, define the set of terminal symbols in γ , denoted by $\text{alph}(\gamma) = B$, and the set of variables of γ , denoted by $\text{var}(\gamma) = Y$. For any symbol $t \in \Sigma \cup \mathcal{X}$ and $\alpha \in PAT_\Sigma$, $|\alpha|_t$ denotes the number of occurrences of t in α .

A substitution (on the variables of α) is a mapping $h : \text{var}(\alpha) \rightarrow \Sigma^*$. For every $x \in \text{var}(\alpha)$, we say that x is substituted by $h(x)$ and $h(\alpha)$ denotes the word obtained by substituting every occurrence of a variable x in α by $h(x)$ and leaving all the terminals unchanged. We say that the pattern α matches a word $w \in \Sigma^+$, if there exists a substitution $h : \text{var}(\alpha) \rightarrow \Sigma^*$ such that $h(\alpha) = w$. The Matching Problem is defined for any family of patterns $P \subseteq PAT$:

Exact Matching Problem for P : `MatchP`

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w , with $|w| = n$.

Question: Is there a substitution h with $h(\alpha) = w$?

In this paper, we will consider an extension of the Matching Problem, in which we allow mismatches between the image of the pattern under a substitution and the matched word.

For words $w_1, w_2 \in \Sigma^*$ with $|w_1| = |w_2|$, the Hamming distance between w_1 and w_2 is defined as $d_{\text{HAM}}(w_1, w_2) = |\{i \mid w_1[i] \neq w_2[i] \wedge 1 \leq i \leq |w_1|\}|$. The Hamming distance describes, therefore, the number of mismatches between two words. For a pattern α and a word w , we can define the Hamming distance between α and w as $d_{\text{HAM}}(\alpha, w) = \min\{d_{\text{HAM}}(h(\alpha), w) \mid h \text{ is a substitution of the variables of } \alpha\}$. With these definitions we can introduce two new pattern matching problems for families of patterns $P \subseteq PAT$. In the first problem, we allow

for a certain distance Δ between the image $h(\alpha)$ of α under a substitution h and the target word w instead of searching for an exact matching. In the second problem, we are interested in finding the substitution h such that the number of mismatches between $h(\alpha)$ and the target word w is minimal, over all possible choices of h .

Approximate Matching Decision Problem for P : **MisMatch $_P$**

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w , with $|w| = n$, an integer $\Delta \leq m$.

Question: Is $d_{\text{HAM}}(\alpha, w) \leq \Delta$?

Approximate Matching Minimisation Problem for P : **MinMisMatch $_P$**

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w , with $|w| = n$.

Question: Compute $d_{\text{HAM}}(\alpha, w)$.

When analysing the number of mismatches between $h(\alpha)$ and w we need to argue about the number of mismatches between corresponding factors of $h(\alpha)$ and w , i.e., the factors occurring between the same positions i and j in both words. To simplify the presentations, for a substitution h that maps a pattern α to a word of the same length as w , we will call the factors $h(\alpha)[i : j]$ and $w[i : j]$ aligned under h . We omit h when it is clear from the context. Moreover, saying that we align a factor $\alpha[i : j]$ to a factor $w[i' : j']$ with a minimal number of mismatches, we mean that we are looking for a substitution h such that $|h(\alpha)| = |w|$, $h(\alpha[i : j])$ is aligned to $w[i' : j']$ under h , and the resulting number of mismatches between $h(\alpha[i : j])$ and $w[i' : j']$ is minimal w.r.t. all other choices for the substitution h .

We make some preliminary remarks. Firstly, in all the problems we consider here, we can assume that the pattern α starts and ends with variables, i.e., $\alpha = x\alpha'y$, with α' pattern and x and y variables. Indeed, if this would not be the case, we could simply reduce the problems by considering them for inputs α' and the word w' obtained by removing from w the prefix and suffix aligned, respectively, to the maximal prefix of α which contains only terminals and the maximal suffix of α which contains only terminals. Clearly, in the case of the exact-matching problem the respective prefixes (suffixes) of w and α must match exactly, while in the case of the approximate-matching problems one needs to account for the mismatches created by these prefixes and suffixes. So, from now on, we will work under the assumption that the patterns we try to align to words start and end with variables.

Secondly, solving **Match $_P$** is equivalent to solving **MisMatch $_P$** for $\Delta = 0$. Also, in a general framework, **MinMisMatch $_P$** can be solved by combining the solution of the decision problem **MisMatch $_P$** with a binary search on the value of Δ . Given that the distance between α and w is at most $n = |w|$, one needs to use the solution for **MisMatch $_P$** a maximum of $\log n$ times in order to find the exact distance between α and w . Sometimes this can be done even more efficiently, as shown in Theorem 3.4. On the other hand, solving **MinMisMatch $_P$** leads directly to a solution for **MisMatch $_P$** .

3 Matching Regular Patterns with Mismatches

A pattern α is *regular* if $\alpha = w_0 \prod_{i=1}^M (x_i w_i)$, with $w_i \in \Sigma^*$. The class of regular patterns is denoted by **Reg**. For example, the pattern $\alpha_0 = \text{abxabyzbaab}$, with $\text{var}\alpha = \{x, y, z\}$ is in **Reg**.

In this section we consider **MisMatch $_{\text{Reg}}$** and **MinMisMatch $_{\text{Reg}}$** .

As mentioned already, a solution for **MisMatch $_{\text{Reg}}$** with distance $\Delta = 0$ is a solution to **Match $_{\text{Reg}}$** . The latter problem can be solved in $\mathcal{O}(n)$ by a greedy approach. As noted in Section 2, we can assume that $w_0 = w_M = \varepsilon$, so $\alpha = (\prod_{i=1}^{M-1} x_i w_i) x_M$. Thus, we identify the last occurrence $w[\ell + 1 : \ell + |w_{M-1}|]$ of w_{M-1} in w , assign the string $w[\ell + |w_{M-1}| + 1 : n]$ to x_M , and then recursively match the pattern $\alpha = (\prod_{i=1}^{M-2} x_i w_i) x_{M-1}$ to $w[1 : \ell]$.

In the following, we propose a solution for $\text{MinMisMatch}_{\text{Reg}}$ which generalizes this approach. Further, we will show a matching lower bound for any algorithm solving $\text{MinMisMatch}_{\text{Reg}}$.

An equivalent formulation of $\text{MinMisMatch}_{\text{Reg}}$ is to find factors $w[\ell_i + 1 : \ell_i + |w_i|]$, with $1 \leq i \leq M-1$, such that $\sum_{i=1}^{M-1} d_{\text{HAM}}(w_i, w[\ell_i + 1 : \ell_i + |w_i|])$ is minimal and $\ell_i + |w_i| + 1 \leq \ell_{i+1}$, for all $i \in \{1, \dots, M-2\}$. In other words, we want to find the $M-1$ factors $w[\ell_i + 1 : \ell_i + |w_i|]$, with i from 1 to $M-1$, such that these factors occur one after the other without overlapping in w , they correspond (in order, from left to right) to the words w_i , for i from 1 to $M-1$, and the total sum of mismatches between $w[\ell_i + 1 : \ell_i + |w_i|]$ and w_i , added up for i from 1 to $M-1$, is minimal.

To approach this problem we need the following data-structures-preliminaries.

Given a word w , of length n , we can construct in $O(n)$ -time *longest common suffix*-data structures which allow us to return in $O(1)$ -time the value $LCS_w(i, j) = \max\{|v| \mid v \text{ is a suffix of both } w[1 : i] \text{ and } w[1 : j]\}$. See [33, 34] and the references therein. Given a word w , of length n , and a word u , of length m , we can construct in $O(n+m)$ -time data structures which allow us to return in $O(1)$ -time the value $LCS_{w,u}(i, j) = \max\{|v| \mid v \text{ is a suffix of both } w[1 : i] \text{ and } u[1 : j]\}$. This is achieved by constructing LCS_w -data structures for wu , as above, and noting that $LCS_{w,u}(i, j) = \min(LCS_w(i, n+j), j)$.

The following two lemmas are based on the data structures defined above and the technique called kangaroo-jump [35]. Their respective proofs can be found in the Appendix B.

► **Lemma 3.1.** *Let w and u , with $|w| = |u| = n$, be two words and δ a non-negative integer. Assume that, in a preprocessing phase, we have constructed $LCS_{w,u}$ -data structures. We can compute $\min(\delta + 1, d_{\text{HAM}}(u, w))$ using $\delta + 1$ $LCS_{w,u}$ queries, so in $O(\delta)$ time.*

► **Lemma 3.2.** *Given a word w , with $|w| = n$, a word u , with $|u| = m < n$, and a non-negative integer δ , we can compute in $O(n\delta)$ time the array $D[m : n]$ with $n - m + 1$ elements, where $D[i] = \min(\delta + 1, d_{\text{HAM}}(w[i - m + 1 : i], u))$.*

The following result is the main technical tool of this section.

► **Theorem 3.3.** *$\text{MisMatch}_{\text{Reg}}$ can be solved in $O(n\Delta)$ time. For an accepted instance w, α, Δ of $\text{MisMatch}_{\text{Reg}}$ we also compute $d_{\text{HAM}}(\alpha, w)$ (which is upper bounded by Δ).*

Proof. Assume $\alpha = \prod_{i=1}^{M-1} (x_i w_i) x_M$ and let $\alpha_\ell = \prod_{i=\ell}^{M-1} (x_i w_i) x_M$, for $\ell \in \{1, \dots, M-1\}$.

A first observation is that the problem can be solved in a standard way by dynamic programming in $O(nm)$ time.

We only give the main idea behind this approach. We can compute the minimum number of mismatches $T[i][j]$ which can be obtained when aligning the suffix of length i of w to the suffix of length j of α , for all $i \leq n$ and $j \leq m$. Clearly, $T[i][j]$ can be computed based on the values $T[i+1][j+1]$ and, if $\alpha[j]$ is a variable, $T[i+1][j]$. The full technicalities of this standard approach are easy to obtain so we do not go into further details.

We present a more efficient approach below.

Our efficient algorithm starts with a preprocessing phase, in which we compute $LCS_{w,u}$ -data structures, where $u = \prod_{i=\ell}^{M-1} w_i$. This allows us to retrieve in constant time answers to LCS_{w,w_i} -queries, for $1 \leq i \leq M-1$.

In the main phase of our algorithm, we compute an $(M-1) \times \Delta$ matrix $\text{Suf}[\cdot][\cdot]$, where, for $\ell \leq M-1$ and $d \leq \Delta$, we have $\text{Suf}[\ell][d] = g$ if and only if $w[g..n]$ is the shortest suffix of w with $d_{\text{HAM}}(\alpha_\ell, w[g : n]) \leq d$.

Once more, we note that the elements of $\text{Suf}[\cdot][\cdot]$ can be computed by a relatively straightforward dynamic programming approach in $O(nM\Delta)$ time. But, the strategy we present here is more efficient than that.

In our algorithm, we first use Lemma 3.2 to compute $Suf[M-1][\cdot]$ in $O(n\Delta)$ time. We simply run the algorithm of that lemma on the input strings w and w_{M-1} and the integer Δ . We obtain an array $D[\cdot]$, where $D[i] = \min(\Delta + 1, d_{\text{HAM}}(w[i - |w_{M-1}| + 1 : i], w_{M-1}))$. We now go with j from $|w_{M-1}|$ to n and, if $D[j] \leq \Delta$, we set $Suf[M-1][D[j]] = j - |w_{M-1}| + 1$. It is clear that $h = Suf[M-1][d]$ will be the starting position of the shortest suffix $w[h : n]$ of w such that $d_{\text{HAM}}(w_{M-1}x_M, w[h : n]) \leq d$. Thus, $Suf[M-1][\cdot]$ was correctly computed, and the time needed to do so is $O(n\Delta)$.

Further, we describe how to compute $Suf[\ell][\cdot]$ efficiently, based on $Suf[\ell+1][\cdot]$ (for ℓ from $M-2$ down to 1). We use the following approach. We go through the positions i of w from right to left and maintain a queue Q . When i is considered, Q stores all elements d such that $Suf[\ell][d]$ was not computed yet until reaching that position, but $i < Suf[\ell+1][d]$. Accordingly, the fact that d is in Q means that with a suitable alignment of w_ℓ ending on position i , we could actually find an alignment with $\leq d$ mismatches of α_ℓ with $w[i - |w_\ell| + 1 : n]$: when Q contains $d, \dots, d-t$, for some $t \geq 0$, an alignment of w_ℓ to $w[i - |w_\ell| + 1 : i]$ with $\leq t$ mismatches would lead to an alignment of α_ℓ with $w[i - |w_\ell| + 1 : n]$ with $\leq d$ mismatches by extending the alignment of $\alpha_{\ell+1}$ to $w[Suf[\ell+1][d-t] : n]$. The values d present in Q at some point are ordered increasingly (the older values are larger), the array $Suf[\ell+1][\cdot]$ is also monotonically increasing, and, as $Suf[\ell][d]$ cannot be set before $Suf[\ell][d']$, for any d and d' such that $d' < d$, the queue Q is actually an interval of integers $[new : old]$, where new is the newest element of Q , and old the oldest one. When we consider position i of the word, if the alignment of w_ℓ ending on position i causes t mismatches, then to be able to set a value $Suf[\ell][d]$, with $d \in Q$, we need to have that $Suf[\ell+1][d-t] > i$. As $Suf[\ell+1][d] > Suf[\ell+1][d-t]$ and $d \in Q$, this means that $d-t \in Q$, so the number of mismatches t must be strictly upper bounded by $|Q|$, in order to be useful. Accordingly, when considering position i , we compute the number $t \leftarrow \min\{d_{\text{HAM}}(w_\ell, w[i - |w_\ell| + 1 : i]), |Q|\}$, and if $t < |Q|$ we set $Suf[\ell][d] \leftarrow i - |w_\ell| + 1$ for all d such that $d-t \in Q$; we also eliminate all these elements d from the queue. Before considering a new position i , we check if $i = Suf[\ell+1][new-1]$, and, if yes, we insert $new-1$ in Q and update $new \leftarrow new-1$.

This computation of $Suf[\ell][\cdot]$ is implemented in the following algorithm:

1. Initialization: We maintain a queue Q , which initially contains only the Δ .
Let $new \leftarrow \Delta$ (this is the top element of the queue).
2. Iteration: For $i = Suf[\ell+1][\Delta] - 1$ down to $|w_\ell|$ we execute the steps a, b, and c:
 - a. Using Lemma 3.1 we compute $t \leftarrow \min(d_{\text{HAM}}(w_\ell, w[i - |w_\ell| + 1 : i]), |Q|)$.
 - b. If $t < |Q|$, we remove from Q all elements d , such that $d-t \geq new$, and set, for each of them, $Suf[\ell][d] \leftarrow i - |w_\ell| + 1$.
 - c. If $Suf[\ell+1][top-1] = i$ then we insert $top-1$ in Q and $top \leftarrow top-1$. Else, if $Suf[\ell+1][top-1] = 0$ then set $i \leftarrow 0$ and exit the loop.
3. Filling-in the remaining positions: Set all the positions of $Suf[\ell][\cdot]$ which were not filled during the above while-loop to 0.

The matrix $Suf[\cdot][\cdot]$ is computed correctly by the above algorithm, as it can be shown by the following inductive argument.

To show that $Suf[\ell][\cdot]$ is computed correctly by our algorithm, under the assumption that $Suf[\ell+1][\cdot]$ was correctly computed, we make several observations.

Firstly, it is clear that $Suf[\ell+1][d] \leq Suf[\ell+1][d+1]$. Secondly, when computed correctly, $Suf[\ell][d]$ should be the rightmost position g of w such that $d_{\text{HAM}}(w[g : n], w_\ell) = t \leq d$ and $Suf[\ell+1][d-t] \geq g + |w_\ell|$. Clearly, if $Suf[\ell][d+1] \neq 0$, then $Suf[\ell][d] < Suf[\ell][d+1]$.

Regarding the algorithm described in the main part of the paper, it is important to observe that the queue Q is ordered increasingly (i.e., the newer is an element in Q , the smaller it is) and the elements of Q form an interval $[new : old]$.

Now, let us show the correctness of the algorithm.

Let d be a non-negative integer, $d \leq \Delta$. Assume that our algorithm sets $Suf[\ell][d] = g$, with $g > 0$.

This means that d was removed from the queue in step 2.b when the for-loop was executed for $i = g + |w_\ell| - 1$. The reason for this removal was that $d_{\text{HAM}}(w[g : g + |w_\ell| - 1], w_\ell) = t \leq |Q| - 1$. Hence, in this step we have removed exactly those elements δ such that $new \leq \delta - t$. Accordingly, we also have that $new \leq d - t$ holds. Let $g' = Suf[\ell + 1][new]$. We thus have $g' > i = g + |w_\ell| - 1$, $d_{\text{HAM}}(\alpha_{\ell+1}, w[g' : n]) \leq new$, and $d_{\text{HAM}}(w_\ell x_\ell, w[g : g' - 1]) = t$. Putting this all together, we get that $d_{\text{HAM}}(\alpha_\ell, w[g : n]) \leq new + t \leq d$.

Now, assume for the sake of a contradiction, that there exists $g'' > g$ such that $d_{\text{HAM}}(\alpha_\ell, w[g'' : n]) \leq d$, i.e., $w[g : n]$ is not the shortest suffix s of w such that $d_{\text{HAM}}(\alpha_\ell, s) \leq d$. In this case, there exists d'' such that $g'' + |w_\ell| - 1 < Suf[\ell + 1][d'']$ and $d'' + d_{\text{HAM}}(w[g'' : g'' + |w_\ell| - 1], w_\ell) \leq d$. Because d is in Q when $i = g + |w_\ell| - 1$ is reached in the for-loop, then d must also be in Q when $i'' = g'' + |w_\ell| - 1$ is reached in the for-loop, because $i < i'' < Suf[\ell + 1][d''] \leq Suf[\ell + 1][d]$. In fact, as $Suf[\ell + 1][d] \geq Suf[\ell + 1][d''] > i''$, it follows that d'' must also be in Q when i'' is reached. Thus, $g \geq d - d''$ and, as we have seen above, $d - d'' \geq d_{\text{HAM}}(w[g'' : g'' + |w_\ell| - 1], w_\ell)$. Moreover, if new'' is the element on the top of the queue when i'' is reached, we have that $new'' \leq d''$. Hence, $new'' + d_{\text{HAM}}(w[g'' : g'' + |w_\ell| - 1], w_\ell) \leq d'' + d_{\text{HAM}}(w[g'' : g'' + |w_\ell| - 1], w_\ell) \leq d$. Therefore, when i'' was reached, all the conditions needed to remove d from Q and set $Suf[\ell][d] \leftarrow g''$ were met. We have reached a contradiction with our assumption that $g'' > g$.

In conclusion, if our algorithm sets $Suf[\ell][d] = g$, with $g > 0$, then $w[g : n]$ is the shortest suffix of w such that $d_{\text{HAM}}(w[g : n], w_\ell) \leq d$. By an analogous argument as the one used above in our proof by contradiction, we can show that if our algorithm sets $Suf[\ell][d] = 0$ then there does not exist any suffix $w[g : n]$ of w such that $d_{\text{HAM}}(w[g : n], w_\ell) \leq d$.

This means that our algorithm computing $Suf[\cdot][\cdot]$ is correct.

To finalize the proof of the theorem, we note that, after computing the entire matrix $Suf[\cdot][\cdot]$, we can accept the instance w, α, Δ of $\text{MisMatch}_{\text{Reg}}$ if and only if there exists $d \leq \Delta$ such that $Suf[1][d] \neq 0$. Moreover, $d_{\text{HAM}}(\alpha, w) = \min(\{d \mid Suf[1][d] \neq 0\} \cup \{+\infty\})$.

In the following we show that this algorithm works in $O(n\Delta)$ time. We will compute the complexity of this algorithm using amortized analysis. Firstly, we observe that the complexity of the algorithm is proportional to the total number of LCS_{w, w_ℓ} -queries we compute in step 2.a, for each $\ell \leq M$ or, in other words, over all executions of the algorithm. Now, we observe that when position i of w is considered (for a certain ℓ), we do $|Q|$ many LCS_{w, w_ℓ} -queries. So, this means that we do one query per each current element of Q (and none if $|Q| = 0$). Thus, the number of queries corresponding to each pair (ℓ, d) which appears in Q at some point equals the number of positions considered between the step when it was inserted in Q and the step when it was removed from Q . This means $O(Suf[\ell + 1][d] - Suf[\ell][d])$ queries corresponding to (ℓ, d) . Summing this up for a fixed d and ℓ from 1 to $M - 2$ we obtain that the overall number of queries corresponding to a fixed δ is $O(Suf[M - 1][d]) = O(n)$. Adding this up for all $d \leq \Delta$, we obtain that the number of LCS -queries performed in our algorithm is $O(n\Delta)$. So, together with the complexity of the initialization of $Suf[M - 1][\cdot]$, the complexity of this algorithm is $O(n\Delta)$.

This algorithm outperforms the other two algorithms solving $\text{MinMisMatch}_{\text{Reg}}$ which we mentioned, and, for $\Delta = 0$, it is a reformulation of the greedy algorithm solving $\text{Match}_{\text{Reg}}$. ◀

► **Theorem 3.4.** $\text{MinMisMatch}_{\text{Reg}}$ can be solved in $O(n\Phi)$ time, where $\Phi = d_{\text{HAM}}(\alpha, w)$.

Proof. We use the algorithm of Theorem 3.3 for $\Delta = 2^i$, for increasing values of i starting with 1 and repeating until the algorithm returns a positive answer and computes $\Phi = d_{\text{HAM}}(\alpha, w)$. The algorithm is clearly correct. Moreover, the value of i which was considered last is such that $2^{i-1} < \Phi \leq 2^i$. So $i = \lceil \log_2 \Phi \rceil$, and the total complexity of our algorithm is $O(n \sum_{i=1}^{\lceil \log_2 \Phi \rceil} 2^i) = O(n\Phi)$. ◀

In order to show that $\text{MinMismatch}_{\text{Reg}}$ and $\text{Mismatch}_{\text{Reg}}$ cannot be solved by algorithms running polynomially faster than the algorithms from Theorems 3.3 and 3.4, we will reduce the Orthogonal Vectors problem OV [10] to $\text{Mismatch}_{\text{Reg}}$. The overall structure of our reduction is similar to the one used for establishing hardness of computing edit distance [5, 11] or LCS [12], however we needed to construct gadgets specific to our problem. We recall the OV problem.

Orthogonal Vectors: OV

Input: Two sets U, V consisting each of n vectors from $\{0, 1\}^d$, where $d \in \omega(\log n)$.

Question: Do vectors $u \in U, v \in V$ exist, such that u and v are orthogonal, i.e., for all $1 \leq k \leq d$, $v[k]u[k] = 0$ holds?

In general, for a vector $u = (u[1], \dots, u[d]) \in \{0, 1\}^d$, the bits $u[i]$ are called coordinates. It is clear that, for input sets U and V as in the above definition, one can solve OV trivially in $\mathcal{O}(n^2d)$ time. The following conditional lower bound is known for OV .

► **Lemma 3.5 (OV-Conjecture).** *OV can not be solved in $\mathcal{O}(n^{2-\epsilon}d^c)$ for any $\epsilon > 0$ and constant c , unless the Strong Exponential Time Hypothesis (SETH) fails.*

See [10, 48] and the references therein for a detailed discussion regarding conditional lower bounds related to OV . In this context, we can show the following result.

► **Theorem 3.6.** *$\text{Mismatch}_{\text{Reg}}$ can not be solved in $\mathcal{O}(|w|^h \Delta^g)$ time (or in $\mathcal{O}(|w|^h |\alpha|^g)$ time) with $h + g = 2 - \epsilon$ for some $\epsilon > 0$, unless the OV -Conjecture fails.*

Proof. We reduce OV to $\text{MinMismatch}_{\text{Reg}}$. For this, we consider an instance of OV : $U = \{u_1, \dots, u_n\}$ and $V = \{v_1, \dots, v_n\}$, with $U, V \subset \{0, 1\}^d$. We transform this OV -instance into a $\text{Mismatch}_{\text{Reg}}$ -instance (α, w, Δ) , where $\Delta = n(d+1) - 1$. More precisely, we ensure that for the respective $\text{Mismatch}_{\text{Reg}}$ -instance, there exists a way to replace the variables with strings leading to exactly $n(d+1)$ mismatches between the image of α and w if and only if no two vectors u_i and v_j are orthogonal. But, if there exists at least one orthogonal pair of vectors u_i and v_j , there also exists a way to replace the variables of α such that the resulting string has strictly less than $n(d+1)$ mismatches to w . Both $|w|$ and $|\alpha|$ are in $\mathcal{O}(nd)$, and can be built in $O(nd)$ time. The reduction consists of three main steps. First we will present a gadget for encoding the single coordinates of vectors u_i and v_i from U and V , respectively. Then we will show another gadget to encode a full vector of each respective set. And, finally, we will show how to assemble these gadgets of the vectors from set U into the word w and from V into α .

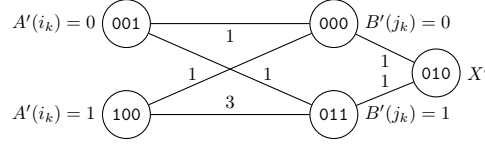
First gadget. Let $u_i = (u_i[1], u_i[2], \dots, u_i[d]) \in U, v_j = (v_j[1], v_j[2], \dots, v_j[d]) \in V$ and let k be a position of these vectors. We define the following gadgets:

$$A'(i_k) = \begin{cases} 001, & \text{if } u_i[k] = 0. \\ 100, & \text{if } u_i[k] = 1. \end{cases} \quad B'(j_k) = \begin{cases} 000, & \text{if } v_j[k] = 0. \\ 011, & \text{if } v_j[k] = 1. \end{cases}$$

Note that, when aligned, the pair of strings $(A'(i_k), B'(j_k))$ produces exactly one mismatch if and only if $u_i[k] \cdot v_j[k] = 0$; otherwise it produces three mismatches. So, $A'(i_k)$ and $B'(j_k)$ encode the single coordinates of u_i and v_j respectively.

48:10 Matching Patterns with Variables Under Hamming Distance

Further, we construct a gadget $X' = 010$ that produces always one mismatch if aligned to any of the strings $B'(j_k)$ corresponding to coordinates $v_j[k]$. See also Figure 1.



■ **Figure 1** Gadgets for the encoding of single coordinates of the vectors. On each edge we wrote the number of mismatches between the strings in the nodes connected by that edge.

Second gadget. The gadget $A(i)$ encodes the vector u_i , for $1 \leq i \leq n$, while the gadget $B(j)$ encodes the vector v_j , for $1 \leq j \leq n$. We construct these gadgets such that aligning $B(j)$ to $A(i)$ with a minimum number of mismatches yields exactly d mismatches, if the two corresponding vectors are orthogonal, and exactly $d + 1$ mismatches, otherwise. Moreover, we show that any other alignment of the gadgets $B(j)$ with other factors of w yields more mismatches.

In order to assemble the gadgets $A(i)$ and $B(j)$, for $1 \leq i, j \leq n$, we extend the terminal alphabet by three new symbols $\{\mathbf{a}, \mathbf{b}, \#\}$, as well as use two fresh variables x_j, y_j for each vector v_j . The gadgets $A(i)$, for all i , and, respectively, the gadgets $B(j)$, for all j , consist of the concatenation of the coordinate gadgets $A'(i_k)$ and, respectively, $B'(j_k)$ from left to right, in ascending order of k . Each two such consecutive gadgets $A'(i_k)$ and $A'(i_{k+1})$ (respectively, $B'(j_k)$ and $B'(j_{k+1})$) are separated by $\#\#\#$. We prepend to $A(i)$ the string \mathbf{bba} and append the string $\mathbf{bbb}X$, where $X = (X'\#\#\#)^{d-1}X'$. In the case of $B(j)$, we prepend $x_j\mathbf{bba}$ and append y_j . The full gadgets $A(i)$ and $B(j)$ are defined as follows.

- $A(i) = \mathbf{bba}A'(i_1)\#\#\#A'(i_2)\#\#\# \dots A'(i_d)\mathbf{bbb}X$
- $B(j) = x_j\mathbf{bba}B'(j_1)\#\#\#B'(j_2)\#\#\# \dots B'(j_d)y_j$.

For simplicity of the exposure, let $B'(j) = \mathbf{bba}B'(j_1)\#\#\#B'(j_2)\#\#\# \dots \#\#\#B'(j_d)$.

Note that $|A(i)|$ is the same for all i , so we can define $M = |A(i)|$.

Final assemblage. To define the word w , we use a new terminal $\$$. The word w is:

- $w = \$^M A(1)\$^M A(2)\$^M \dots A(n)\$^M A(1)\$^M A(2) \dots \$^M A(n)\M

To define α , we use two new fresh variables x and y . The pattern α is:

- $\alpha = x\$^M B(1)\$^M B(2)\$^M \dots \$^M B(n)\$^M y$.

The correctness of the reduction. We show that there exists a way to align α with w with $< n(d + 1)$ mismatches if and only if a pair of orthogonal vectors $u_i \in U$ and $v_j \in V$ exists. Otherwise, there exists an alignment of α to w with exactly $n(d + 1)$ mismatches.

To formally prove that the reduction fulfills this requirement, we proceed as follows.

A general idea: the repetition of the gadgets $A(i)$ in the word w guarantees that, if needed, a pair of gadgets $A(i)$ and $B(j)$, corresponding to the vectors $u_i \in U$ and, respectively, $v_j \in V$, can be aligned. More precisely, we can align $B'(j)$ to $\mathbf{bba}A'(i_1)\#\#\# \dots A'(i_d)$. The variables x, y and x_j, y_j , for $j \in \{1, \dots, n\}$, act as spacers: they allow us to align a string $B'(j)$ to the desired factor of w . This kind of alignment is enough for our purposes, as we only need to find one orthogonal pair of vectors, not all of them; however, we need enough space in w for the factors of α occurring before and after $B'(j)$, thus the repetition of the $A(i)$ gadgets.

We now analyse how a factor $B'(j)$ can be aligned to a factor of w . The main idea is to show that if there are no orthogonal vectors, then any alignment of $B'(j)$ to a factor of w creates at least $d + 1$ mismatches. Otherwise, we can align it with d mismatches only.

Case 1: $B'(j)$ is aligned to a factor $w[i : h]$ of w which starts with \$. Then the prefix **ba** of $B'(j)$ causes at least two mismatches, as the first **b** in **ba** is aligned to a \$ letter, while the **a** is aligned to either a **b** letter (from a **ba** factor) or a \$ letter. The rest of $B'(j)$ causes, overall, at least d mismatches, one per each group $B'(j_k)$. So, in this case, we have at least $d + 2$ mismatches caused by $B'(j)$.

Case 2: $B'(j)$ is aligned a factor $w[i : h]$ of w which ends with \$. Then, its prefix **ba** cannot be aligned to a factor **ba** of w . So, the **a** of the prefix **ba** of $B'(j)$ produces one mismatch, while the suffix $B'(j_d)$ causes at least 2 mismatches. The rest of $B'(j)$ causes at least $d - 1$ mismatches, one per each remaining group $B'(j_k)$. So, in this case, we have again at least $d + 2$ mismatches caused by $B'(j)$.

Case 3: $B'(j)$ is aligned exactly to the factor **ba** $A'(i_1)$ ###... $A'(i_d)$ and u_i and v_j are orthogonal, then $B'(j)$ causes exactly d mismatches.

Case 4: $B'(j)$ is aligned exactly to the factor **ba** $A'(i_1)$ ###... $A'(i_d)$ and u_i and v_j are not orthogonal, then $B'(j)$ causes at least $d + 2$ mismatches.

Case 5: $B'(j)$ is aligned exactly to the factor **bb** X , then $B'(j)$ causes $d + 1$ mismatches.

Case 6: $B'(j)$ is aligned to a factor starting strictly inside **ba** $A'(i_1)$ ###... $A'(i_d)$, then the prefix **ba** of $B'(j)$ cannot be aligned to a factor **ba** of w , so it causes at least two mismatches (from the alignment of **ba**). The rest of $B'(j)$ causes at least d mismatches, one per each group $B'(j_k)$. So, overall, $B'(j)$ causes at least $d + 2$ mismatches in this case.

To ease the understanding, cases 3 and 4 are illustrated in the following table: when aligning $A(i)$ to $B(j)$, to obtain the desired number of mismatches, we can match the parts of $A(i)$ to the parts of $B(j)$ as described in this table in the two cases 3. and 4.

Gadget	I	II	III	IV	mismatches
$A(i) =$	ε	ba $A'(i_1)$ ###...### $A'(i_d)$	bb X' ###...### X'	ε	
3. $B(j) =$	x_j	ba $B'(j_1)$ ###...### $B'(j_d)$	y_j	ε	d (in II)
4. $B(j) =$	ε	x_j	ba $B'(j_1)$ ###...### $B'(j_d)$	y_j	$d + 1$ (in IV)

Wrapping up, there are no other ways than those described in cases 1-6 above in which $B'(j)$ can be aligned to a factor of w . In particular, in order to reach an alignment with at most $n(d + 1) - 1$ mismatches, at least one $B'(j)$ should be aligned to a factor of w such that it only causes d mismatches (as in case 3). Thus, in that case we would have a pair of orthogonal vectors. Conversely, if there exist u_i and v_j which are orthogonal and $i \geq j$, then we can align $B'(j)$ to the occurrence of **ba** $A'(i_1)$ ###... $A'(i_d)$ from the first $A(i)$ and all the other gadgets $B'(\ell)$ to factors **bb** X , and obtain a number of $n(d + 1) - 1$ mismatches. Note that such an alignment is possible as there exist at least $j - 1$ factors **bb** X before the first $A(i)$ and at least n more occurrences of **bb** X after it; moreover the variables x_ℓ and y_ℓ can be used to align as desired the strings $B'(v_\ell)$ to the respective **bb** X factors of w . If there exist u_i and v_j which are orthogonal and $i < j$, then we can align $B'(j)$ to the occurrence of **ba** $A'(i_1)$ ### $A'(i_2)$ ###... $A'(i_d)$ from the second $A(i)$ and all the other gadgets $B'(\ell)$ to factors **bb** X , and obtain again a number of $n(d + 1) - 1$ mismatches. This is possible for similar reasons to the ones described above.

This shows that our reduction is correct. The instance of **OV** defined by U and V contains two orthogonal vectors if and only the instance of **MisMatch_{Reg}** defined by w, α , and $\Delta = n(d + 1) - 1$ can be answered positively. Moreover, the instance of **MisMatch_{Reg}** can be constructed in $O(nd)$ time and we have that $|w|, |\alpha|, \Delta \in \Theta(nd)$.

Assume now that there exists a solution of $\text{MisMatch}_{\text{Reg}}$ running in $O(|w|^g|\alpha|^h)$ with $g+h=2-\epsilon$ for some $\epsilon < 0$. This would lead to a solution for OV running in $O(nd+(nd)^{2-\epsilon})$, a contradiction to the OV -conjecture. Similarly, if there exists a solution of $\text{MisMatch}_{\text{Reg}}$ running in $O(|w|^g\Delta^h)$ with $g+h=2-\epsilon$ for some $\epsilon < 0$, then there exists a solution for OV running in $O(nd+(nd)^{2-\epsilon})$, a contradiction to the OV -conjecture. This proves our statement. \blacktriangleleft

► **Remark 3.7.** An immediate consequence of the previous theorem is that $\text{MinMisMatch}_{\text{Reg}}$ can not be solved in $\mathcal{O}(n^h d_{\text{HAM}}(\alpha, w)^g)$ time (or in $\mathcal{O}(|w|^h|\alpha|^g)$ time) with $h+g=2-\epsilon$ for some $\epsilon > 0$, unless the OV -Conjecture fails. Thus, as $d_{\text{HAM}}(\alpha, w) \leq |\alpha|$, $\text{MinMisMatch}_{\text{Reg}}$ and $\text{MisMatch}_{\text{Reg}}$ cannot be solved polynomially faster than our algorithms, unless the OV -Conjecture fails.

4 Patterns with Repeated Variables

In Section 3 we have shown that if no variable occurs more than once in the input pattern α , then the problems MisMatch and MinMisMatch can be solved in polynomial time. Let us now consider patterns where variables are allowed to occur more than once, i.e., patterns with repeated variables. Firstly, we recall two measures of the structural complexity of patterns.

For every variable $x \in \text{var}(\alpha)$, the scope of x in α is defined by $\text{sc}_\alpha(x) = [i : j]$, where i is the leftmost and j the rightmost occurrence of x in α . The scopes of the variables $x_1, \dots, x_k \in \text{var}(\alpha)$ coincide in α if $\bigcap_{i=1}^k \text{sc}(x_i) \neq \emptyset$. By $\text{scd}(\alpha)$ we denote the scope coincidence degree of α : the maximum number of variables in α whose scopes coincide. By kSCD we denote the class of patterns whose scope coincidence degree is at most k .

Given a pattern α , with p variables, a marking sequence of α is an ordering $x_1 < x_2 < \dots < x_p$ of $\text{var}(\alpha)$. The skeleton α_{var} of α is obtained from α by removing all the terminals. A marking of α_{var} w.r.t. a marking sequence $x_1 < x_2 < \dots < x_p$ of α is a p -steps procedure: in step i we mark all occurrences of variable x_i . The pattern α is called k -local if and only if there exists a marking sequence of $x_1 < x_2 < \dots < x_p$ of α such that, for i from 1 to p , the variables marked in the first i steps of the marking of α_{var} w.r.t. this marking sequence form at most k non-overlapping length-maximal factors in α_{var} ; the respective marking sequence is called witness for the k -locality of α . By kLOC we denote the class of k -local patterns. See [18, 15] for an extended discussion and examples regarding k -locality.

Several more particular classes which we consider in this context are the following:

- The class of unary patterns 1Var : $\alpha \in \text{1Var}$ if there exists $x \in X$ such that $\text{var}(\alpha) = \{x\}$; example: $\alpha_1 = \text{abxabxxbaab} \in \text{1Var}$.
- The class of one-repeated-variable patterns 1RepVar : $\alpha \in \text{1RepVar}$ if there exists at most one variable $x \in X$ such that $|\alpha|_x > 1$; example: $\alpha_2 = \text{abxyabzxxbaabv} \in \text{1RepVar}$.
- The class $\text{NonCross} = \text{1SCD}$, called the class of non-cross patterns; as examples, consider $\alpha_3 = \text{abxxyabzzbbvvvabvu} \in \text{NonCross} \setminus \text{1RepVar}$ and $\alpha_4 = \text{abxyabzxxbbvabx} \in \text{1RepVar} \setminus \text{NonCross}$. Note that $\alpha \in \text{NonCross}$ if and only if α can be written as the concatenation of several 1Var -patterns, whose variables are pairwise distinct. Thus, NonCross -patterns are 1-local.

Note that in a NonCross -pattern α , for any two variables $x, y \in \text{var}(\alpha)$, where the last occurrence of y is to the right of the first occurrence of x in α , we can actually write $\alpha = \beta x \gamma y \delta$ such that $x, y \notin \text{var}(\gamma)$, $x \notin \text{var}(\delta)$, and $y \notin \text{var}(\beta)$. In other words, there are no interleaved occurrences of two variables. Moreover, if $\alpha \in \text{NonCross}$, then α is 1-local: the marking sequence is obtained by ordering the variables according to the position of their first

occurrence. Clearly, $1\text{Var} \subset 1\text{RepVar}$ and $1\text{Var} \subset \text{NonCross}$, but 1RepVar and NonCross are incomparable. Indeed, if $\alpha \in \text{NonCross}$ then α is 1-local and 1RepVar contains patterns α with $\text{scd}(\alpha) = 2$.

Further, if α is a pattern and $x \in \text{var}(\alpha)$, then an x -block is a factor $\alpha[i : j]$ such that $\alpha[i : j] \in 1\text{Var}$ with $\text{var}(\alpha[i : j]) = x$ and it is length-maximal with this property: it cannot be extended to the right or to the left without introducing a variable different from x .

The next lemma is fundamental for the results of this section.

► **Lemma 4.1.** *Given a set of words $w_1, \dots, w_p \in \Sigma^m$, we can find in $O(|\Sigma| + mp)$ a median string for $\{w_1, \dots, w_p\}$, i.e. a string w such that $\sum_{j=1}^p d_{\text{HAM}}(w_i, w)$ is minimal.*

Proof. We will use an array C with Σ elements, called counters, indexed by the letters of Σ , and all initially set to 0. For each i between 1 and m , we count how many times each letter of Σ occurs in the multi-set $\{w_1[i], w_2[i], \dots, w_p[i]\}$ using C . Let $w[i]$ be the most frequent letter of this multi-set. After computing $w[i]$, we reset the counters which were changed in this iteration, and repeat the algorithm for $i + 1$. After going through all values of i , we return the word $w = w[1]w[2] \dots w[m]$ as the answer to the problem. The correctness of the algorithm is immediate, while its complexity is clearly $O(|\Sigma| + mp)$. ◀

The typical use of this lemma is the following: we identify the factors of w to which a repeated variable is aligned, and then compute the optimal assignment of this variable. Based on this, the following theorem can now be shown. The corresponding proof can be found in the full version of this paper [30].

► **Theorem 4.2.** $\text{MinMismatch}_{1\text{Var}}$ and $\text{Mismatch}_{1\text{Var}}$ can be solved in $O(n)$ time.

By a standard dynamic programming approach, we use the previous result to obtain a polynomial-time solution for $\text{MinMismatch}_{\text{NonCross}}$ based on the solution for $\text{MinMismatch}_{1\text{Var}}$ (in the statement, $p = |\text{var}(\alpha)|$). The corresponding proof can be found in the full version of this paper [30].

► **Theorem 4.3.** $\text{MinMismatch}_{\text{NonCross}}$ and $\text{Mismatch}_{\text{NonCross}}$ can be solved in $O(n^3p)$ time.

The results presented so far show that MinMismatch_P and Mismatch_P can be solved in polynomial time, as long as we do not allow interleaved occurrences of variables in the patterns of the class P . We now consider the case of 1RepVar -patterns, the simplest class of patterns which permits interleaved occurrences of variables. For simplicity, in the results regarding 1RepVar we assume that the variable which occurs more than once in the input pattern is denoted by x . The corresponding proof and the proof of the following more general result can be found in the Appendix B.

► **Theorem 4.4.** $\text{MinMismatch}_{1\text{RepVar}}$ and $\text{Mismatch}_{1\text{RepVar}}$ can be solved in $O(n^{k+2}m)$ time, where k is the number of x -blocks in the input pattern α .

► **Theorem 4.5.** $\text{MinMismatch}_{\text{kLOC}}$ and $\text{Mismatch}_{\text{kLOC}}$ can be solved in $O(n^{2k+2}m)$ time.

Note that NonCross -patterns are 1-local, while the locality of an 1RepVar -pattern is upper bounded by the number of x -blocks. However, the algorithms we obtained in those particular cases are more efficient than the ones which follow from Theorem 4.5.

The fact that Lemma 4.1 is used as the main building block for our results regarding Mismatch_P and MinMismatch_P for $P \in \{1\text{RepVar}, \text{kLOC}\}$, suggests that these problems could be closely related to the following well-studied problem [36, 20, 7, 13].

Consensus Patterns: CP

Input: k strings $w_1, \dots, w_k \in \Sigma^\ell$, integer $m \in \mathbb{N}$ with $m \leq \ell$, an integer $\Delta \leq mk$.

Question: Do the strings s , of length m , and s_1, \dots, s_k , factors of length m of each w_1, \dots, w_k , respectively, exist, such that $\sum_{i=1}^k d_{\text{HAM}}(s_i, s) \leq \Delta$?

Exploiting this connection, and following the ideas of [36], we can show the following theorem. In this theorem we restrict to the case when the input word w of $\text{MinMisMatch}_{1\text{RepVar}}$ is over $\Sigma = \{1, \dots, \sigma\}$ of constant size σ .

► **Theorem 4.6.** *For each constant $r \geq 3$, there exists an algorithm with run-time $O(n^{r+3})$ for $\text{MinMisMatch}_{1\text{RepVar}}$ whose output distance is at most $\min \left\{ 2, \left(1 + \frac{4\sigma-4}{\sqrt{\epsilon}(\sqrt{4r+1}-3)} \right) \right\} d_{\text{HAM}}(\alpha, w)$.*

The proof can be found in the Appendix B. It remains open whether other algorithmic results related to CP (such as those from, e.g., [8, 9, 40]) apply to our setting too.

In the following we show two hardness results which explain why the algorithms in Theorems 4.4 and 4.6 are interesting.

► **Theorem 4.7.** *$\text{MisMatch}_{1\text{RepVar}}$ is $W[1]$ -hard w.r.t. the number of x -blocks.*

Proof. We reduce CP to $\text{MisMatch}_{1\text{RepVar}}$, such that an instance of CP with k different input strings is mapped to an instance of $\text{MisMatch}_{1\text{RepVar}}$ with $k+1$ x -blocks (where x is the repeated variable), each containing exactly one occurrence of x .

Hence, we consider an instance of CP which consists of k strings $w_1, \dots, w_k \in \Sigma^\ell$ of length ℓ and two integer m, Δ defining the length of the target factors and the number of allowed mismatches, respectively.

The instance of $\text{MisMatch}_{1\text{RepVar}}$ which we construct consists of a text w and a pattern α , such that α contains $k+1$ x -blocks, each with exactly one occurrence of x , and is of polynomial size w.r.t. the size of the CP-instance. Moreover, the number of mismatches allowed in this instance of $\text{MisMatch}_{1\text{RepVar}}$ is $\Delta' = m + \Delta$. That is, if there exists a solution for the CP-instance with Δ allowed mismatches, then, and only then, we should be able to find a solution of the $\text{MisMatch}_{1\text{RepVar}}$ -instance with $\Delta + m$ mismatches.

The construction of the $\text{MinMisMatch}_{1\text{RepVar}}$ is realized in such a way that the word w encodes the input strings, while α creates the mechanism for selecting the string s and corresponding factors s_1, \dots, s_k . The general idea is that x should be mapped to s , and the factors to which the occurrences of x are aligned should correspond to the strings s_1, \dots, s_k .

The structure of the word w and that of the pattern α ensure that, in an alignment of α with w which cannot be traced back to a admissible solution for the CP-instance (that is, the occurrences of x are not aligned to factors of length m of the words w_1, \dots, w_k or x is not mapped to a string of length m) we have at least $M \gg \Delta'$ mismatches, hence it cannot lead to a positive answer for the constructed instance of $\text{MisMatch}_{1\text{RepVar}}$.

The reduction consists of three main steps. Firstly, we present a pair of gadgets to encode the relation of the strings w_i and their factors s_i , for i from 1 to k . Then, we present a second pair of gadgets, which ensures that, in a positive solution of $\text{MisMatch}_{1\text{RepVar}}$, the variable x can only be mapped to a string of length m , corresponding to the string s . Finally, we show how to assemble these gadgets into the input word w and the input pattern α for $\text{MisMatch}_{1\text{RepVar}}$.

First pair of gadgets. We introduce the new letters $\{a, b\}$, not contained in the input alphabet of the CP-instance, as well as the variable x and two fresh variables y_i, z_i , for each i from 1 to k . We construct the following two gadgets for each input string w_i with $1 \leq i \leq k$.

- A gadget to be included in w : $g_i = w_i \overbrace{a^M b^M \dots a^M b^M}^M$.
- A gadget to be included in α : $f_i = y_i x z_i \overbrace{a^M b^M \dots a^M b^M}^M$.

These gadgets allows us to align the i^{th} occurrence of x to an arbitrary factor of the word w_i , for i from 1 to k .

Second pair of gadgets. In this case, we use three new letters $\{c, d, \$\}$ which are not contained in the input alphabet of CP. Also, let $M = (k\ell)^2$. We define two new gadgets.

- A gadget to be included in w : $A_w = \overbrace{c^M d^M \dots c^M d^M}^M \m .
- A gadget to be included in α : $A_\alpha = \overbrace{c^M d^M \dots c^M d^M}^M x$.

These gadgets enforce that, in an alignment of α and w , the variable x is mapped to a string of length m , at the cost of exactly m extra mismatches. Note that, because $\Delta \leq km$, we have that $M \gg \Delta$.

Final assemblage. The word w and the pattern α are defined as follows.

- $w = g_1 g_2 \dots g_k A_w$ and $\alpha = f_1 f_2 \dots f_k A_\alpha$.

To wrap up, the instance of $\text{MinMismatch}_{1\text{RepVar}}$ is defined by $w, \alpha, \Delta + m$.

The correctness of the reduction. We will show that our reduction is correct by a detailed case analysis. We consider an alignment of α and w with minimal number of mismatches, and we make the following observations.

- A. Firstly, if every g_i is aligned to f_i , for i from 1 to k , it is immediate that x is mapped to a string of length m , as the last occurrence of x will be aligned to the $\m suffix of w . Thus, the total number of mismatches between α and w in an alignment with a minimum number of mismatches is upper-bounded by $(k+1)m$.
- B. Secondly, we assume, for the sake of a contradiction, that the length of the image of x is not m . If $|x| > m$ (respectively, $|x| < m$) then the prefix $(c^M d^M)^M$ of A_α is aligned to a factor of w which starts strictly to the left of (respectively, to the right of) the first position of the prefix $(c^M d^M)^M$ of A_w . It is not hard to see that this causes at least M mismatches. Indeed, in the case when $|x| > m$, if the factor $(c^M d^M)^M$ of α is aligned to a factor that starts at least M position to the left of the factor $(c^M d^M)^M$ of w , the conclusion is immediate; if the factor $(c^M d^M)^M$ starts less than M positions to the left of the factor $(c^M d^M)^M$ of w , then each group c^M in α will be aligned to a factor of w that includes at least a d letter, so we again reach the conclusion. In the case when $|x| < m$, then, again, each group c^M in α will be aligned to a factor of w that includes at least a d letter, so the alignment leads to at least M mismatches.

So, we can assume from now on that x is mapped to a string of length m . This also implies that A_α and A_w are aligned, so we will largely neglect them from now on.

- C. Thirdly, we assume that there exists i such that $|h(y_i)| + |h(z_i)| \neq |w_i| - m$. Let $j = \min\{i \leq k \mid |h(y_i)| + |h(z_i)| \neq |w_i| - m\}$. Then the suffixes $(a^M b^M)^M$ of g_j and f_j do not align perfectly to each other. If $|h(y_j)| + |h(z_j)| < |w_j| - m$, then the suffix $(a^M b^M)^M$ of f_j is aligned to a factor of w which starts inside w_j . This immediately causes at least M mismatches, as each group a^M will overlap to a group of which contains at least one b letter. If $|h(y_j)| + |h(z_j)| > |w_j| - m$, then the suffix $(a^M b^M)^M$ of f_j is aligned to a factor of w which starts strictly to the right of the factor w_j . However, because $M = (k\ell)^2 \gg k\ell$,

and f_j and g_j are followed by the same number of factors $(\mathbf{a}^M \mathbf{b}^M)^M$ (until the factors A_α and A_w are reached), the factor corresponding to the suffix $(\mathbf{a}^M \mathbf{b}^M)^M$ of f_j cannot start more than $k\ell$ positions to the right of w_j . It is then immediate that this factor $(\mathbf{a}^M \mathbf{b}^M)^M$ of f_j will cause at least M mismatches: each group \mathbf{a}^M will overlap to a group of which contains at least one \mathbf{b} letter.

So, from now on we can assume that the factors $(\mathbf{a}^M \mathbf{b}^M)^M$ of g_j and f_j are aligned.

- D. At this point, it is clear that in each alignment of α and w which fulfils the conditions described in items B and C: the variable x is mapped to a string of length m , and its first k occurrences are aligned to factors of the words w_1, \dots, w_k . We will now show that for each alignment of α and w in which the image of x contains a $\$$ symbol and fulfils the conditions above, there exists an alignment of α and w with at most the same number of mismatches, in which the image of x does not contain a $\$$ symbol and, once more, fulfils the conditions B and C. Assume that in our original alignment x is mapped to a string u_x of length m such that $u_x[i] = \$$. Let u_1, \dots, u_k be the factors of w_1, \dots, w_k , respectively, to which the first occurrences of the variable x are aligned. Consider the string u'_x which is obtained from u_x by simply replacing the $\$$ symbol on position i by $u_1[i]$. And then consider the alignment of α and w which is obtained from the original alignment by changing the image of x to u'_x instead of u_x . When compared to the original alignment, the new alignment has an additional mismatch caused by the occurrence of x aligned to $\m , but at least one less mismatch caused by the alignments of the first k occurrences of x . Indeed, in the original alignment, the i^{th} position of u_x was a mismatch to the i^{th} position of any string u_1, \dots, u_k , but now at least the i^{th} positions of w_1 and u'_x coincide. This shows that our claim holds. A similar argument shows that for any alignment in which x is mapped to a string containing other letters than the input letters from the CP-instance there exists an alignment in which x is mapped to a string containing only letters from the CP-instance.

Hence, from now on we can assume that the factors $(\mathbf{a}^M \mathbf{b}^M)^M$ of g_j and f_j are aligned and that the image of x has length m and is over the input alphabet of CP-instance.

Based on the observations A-D, we can show that the reduction has the desired properties. If the CP-instance admits a solution s, s_1, \dots, s_k which causes a number of mismatches less or equal to Δ , then we can produce an alignment of α to w as follows. We map x to s and, for i from 1 to k , we map x_i and y_i to the prefix of w_i occurring before s_i and, respectively, the suffix of w_i occurring after s_i . This leads to $\Delta + m$ mismatches between α and w , so the input $(w, \alpha, \Delta + m)$ of $\text{MisMatch}_{1\text{RepVar}}$ is accepted. Conversely, if we have an alignment of α and w with at most $\Delta + m$ mismatches, then we have an alignment with the same number of mismatches which fulfils the conditions summarized at the end of item D above. Hence, we can define s as the image of x in this alignment, and the strings s_1, \dots, s_k as the factors of w aligned to the first k occurrences of x from α . Clearly, for i between 1 and k , s_i is a factor of w_i . As m mismatches of the alignment were caused by the alignment of the last x to $\m , we get that $\sum_{i=1}^k d_{\text{HAM}}(s, s_i) \leq \Delta$. Thus, the instance of CP is accepted.

This concludes the proof of the correctness of our reduction. As M is clearly of polynomial size w.r.t. the size of the CP-instance, it follows that both w and α are of polynomial size $\mathcal{O}(kM^2)$. Therefore, the instance of $\text{MinMisMatch}_{1\text{RepVar}}$ can be computed in polynomial time, and our entire reduction is done in polynomial time. Moreover, we have shown that the instance $(w, \alpha, \Delta + M)$ of $\text{MinMisMatch}_{1\text{RepVar}}$ is answered positively if and only if the original instance of CP is answered positively. Finally, as the number of x blocks in α is $k + 1$, where k is the number of input strings in the instance of CP, and CP is $W[1]$ -hard with respect to this parameter, it follows that $\text{MinMisMatch}_{1\text{RepVar}}$ is also $W[1]$ -hard when the number of k -blocks in α is considered as parameter. This completes our proof. \blacktriangleleft

Note that the pattern α constructed in the reduction above is $k-1$ -local (and not k -local): a witness marking sequence is $z_1 < y_2 < z_2 < y_3 < \dots < z_{k-1} < y_k < x < y_1 < z_k$. Thus, $\text{MisMatch}_{1\text{RepVar}}$ is $W[1]$ -hard w.r.t. locality of the input pattern as well. Also, it is easy to see that $\text{scd}(\alpha) = 2$, and, by the results of [42], this shows that the treewidth of the pattern α , as defined in the same paper, is at most 3. Thus, even for classes of patterns with constant scd , number or repeated variables, or treewidth, the problems MisMatch_P and MinMisMatch_P can become intractable. In Theorem 4.6 we have shown that $\text{MinMisMatch}_{1\text{RepVar}}$ admits a polynomial time approximation scheme (for short, PTAS). We will show in the following that it does not admit an efficient PTAS (for short, EPTAS), unless $FPT = W[1]$. This means that there is no PTAS for $\text{MinMisMatch}_{1\text{RepVar}}$ such that the exponent of the polynomial in its running time is independent of the approximation ratio. To show this, we consider an optimisation variant of the problem CP, denoted minCP . In this problem, for k strings $w_1, \dots, w_k \in \Sigma^\ell$ of length ℓ and an integer $m \in \mathbb{N}$ with $m \leq \ell$, we are interested in the smallest non-negative integer Δ for which there exist strings s , of length m , and s_1, \dots, s_k , factors of length m of each w_1, \dots, w_k , respectively, such that $\sum_{i=1}^k d_{\text{HAM}}(s_i, s) = \Delta$. In [7], it is shown that minCP has no EPTAS unless $FPT = W[1]$. We can use this result and the reduction from the Theorem 4.7 to show the following result (see Appendix B).

► **Theorem 4.8.** $\text{MinMisMatch}_{1\text{RepVar}}$ has no EPTAS unless $FPT = W[1]$.

References

- 1 Amihoud Amir, Moshe Lewenstein, and Ely Porat. Faster algorithms for string matching with k mismatches. *J. Algorithms*, 50(2):257–275, 2004. doi:10.1016/S0196-6774(03)00097-X.
- 2 Amihoud Amir and Igor Nor. Generalized function matching. *J. Discrete Algorithms*, 5:514–523, 2007. doi:10.1016/j.jda.2006.10.001.
- 3 Dana Angluin. Finding patterns common to a set of strings. *J. Comput. Syst. Sci.*, 21(1):46–62, 1980. doi:10.1016/0022-0000(80)90041-0.
- 4 Arturs Backurs and Piotr Indyk. Which regular expression patterns are hard to match? In *Proc. 57th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2016*, pages 457–466, 2016. doi:10.1109/FOCS.2016.56.
- 5 Arturs Backurs and Piotr Indyk. Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). *SIAM J. Comput.*, 47(3):1087–1097, 2018. doi:10.1145/2746539.2746612.
- 6 Philip Bille and Martin Farach-Colton. Fast and compact regular expression matching. *Theor. Comput. Sci.*, 409(3):486–496, 2008. doi:10.1016/j.tcs.2008.08.042.
- 7 Christina Boucher, Christine Lo, and Daniel Lokshantov. Consensus patterns (probably) has no EPTAS. In *Proc. 23rd Annual European Symposium, ESA*, volume 9294 of *Lecture Notes in Computer Science*, pages 239–250, 2015. doi:10.1007/978-3-662-48350-3_21.
- 8 Brona Brejová, Daniel G. Brown, Ian M. Harrower, Alejandro López-Ortiz, and Tomás Vinar. Sharper upper and lower bounds for an approximation scheme for consensus-pattern. In *Proc. 16th Annual Symposium Combinatorial Pattern Matching, CPM 2005*, volume 3537 of *Lecture Notes in Computer Science*, pages 1–10, 2005. doi:10.1007/11496656_1.
- 9 Brona Brejová, Daniel G. Brown, Ian M. Harrower, and Tomás Vinar. New bounds for motif finding in strong instances. In *Proc. 17th Annual Symposium Combinatorial Pattern Matching, CPM 2006*, volume 4009 of *Lecture Notes in Computer Science*, pages 94–105, 2006. doi:10.1007/11780441_10.
- 10 Karl Bringmann. Fine-grained complexity theory (tutorial). In *Proc. 36th International Symposium on Theoretical Aspects of Computer Science, STACS 2019*, volume 126 of *LIPICs*, pages 4:1–4:7, 2019. doi:10.4230/LIPICs.STACS.2019.4.

- 11 Karl Bringmann and Marvin Künnemann. Quadratic conditional lower bounds for string problems and dynamic time warping. In *Proc. 56th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 79–97, 2015. doi:10.1109/FOCS.2015.15.
- 12 Karl Bringmann and Marvin Künnemann. Multivariate fine-grained complexity of longest common subsequence. In *Proc. 29th ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, pages 1216–1235. SIAM, 2018. doi:10.1137/1.9781611975031.79.
- 13 Laurent Bulteau and Markus L. Schmid. Consensus strings with small maximum distance and small distance sum. *Algorithmica*, 82(5):1378–1409, 2020. doi:10.1007/s00453-019-00647-9.
- 14 Cezar Câmpeanu, Kai Salomaa, and Sheng Yu. A formal study of practical regular expressions. *Int. J. Found. Comput. Sci.*, 14:1007–1018, 2003. doi:10.1142/S012905410300214X.
- 15 Katrin Casel, Joel D. Day, Pamela Fleischmann, Tomasz Kociumaka, Florin Manea, and Markus L. Schmid. Graph and string parameters: Connections between pathwidth, cutwidth and the locality number. In *Proc. 46th International Colloquium on Automata, Languages, and Programming, ICALP 2019*, volume 132 of *LIPICs*, pages 109:1–109:16, 2019. doi:10.4230/LIPICs.ICALP.2019.109.
- 16 Panagiotis Charalampopoulos, Tomasz Kociumaka, and Philip Wellnitz. Faster approximate pattern matching: A unified approach. In *Proc. 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 978–989, 2020. doi:10.1109/FOCS46700.2020.00095.
- 17 Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on strings*. Cambridge University Press, 2007. doi:10.1017/CB09780511546853.
- 18 Joel D. Day, Pamela Fleischmann, Florin Manea, and Dirk Nowotka. Local patterns. In *Proc. 37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2017*, volume 93 of *LIPICs*, pages 24:1–24:14, 2017. doi:10.4230/LIPICs.FSTTCS.2017.24.
- 19 Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. Document spanners: A formal approach to information extraction. *J. ACM*, 62(2):12:1–12:51, 2015. doi:10.1145/2699442.
- 20 Michael R. Fellows, Jens Gramm, and Rolf Niedermeier. On the parameterized intractability of motif search problems. *Comb.*, 26(2):141–167, 2006. doi:10.1007/s00493-006-0011-4.
- 21 Henning Fernau, Florin Manea, Robert Mercas, and Markus L. Schmid. Revisiting Shinohara’s algorithm for computing descriptive patterns. *Theor. Comput. Sci.*, 733:44–54, 2018. doi:10.1016/j.tcs.2018.04.035.
- 22 Henning Fernau, Florin Manea, Robert Mercas, and Markus L. Schmid. Pattern matching with variables: Efficient algorithms and complexity results. *ACM Trans. Comput. Theory*, 12(1):6:1–6:37, 2020. doi:10.1145/3369935.
- 23 Henning Fernau and Markus L. Schmid. Pattern matching with variables: A multivariate complexity analysis. *Inf. Comput.*, 242:287–305, 2015. doi:10.1016/j.ic.2015.03.006.
- 24 Henning Fernau, Markus L. Schmid, and Yngve Villanger. On the parameterised complexity of string morphism problems. *Theory Comput. Syst.*, 59(1):24–51, 2016. doi:10.1007/s00224-015-9635-3.
- 25 Dominik D. Freydenberger. Extended regular expressions: Succinctness and decidability. *Theory of Comput. Syst.*, 53:159–193, 2013. doi:10.1007/s00224-012-9389-0.
- 26 Dominik D. Freydenberger. A logic for document spanners. *Theory Comput. Syst.*, 63(7):1679–1754, 2019. doi:10.1007/s00224-018-9874-1.
- 27 Dominik D. Freydenberger and Mario Holldack. Document spanners: From expressive power to decision problems. *Theory Comput. Syst.*, 62(4):854–898, 2018. doi:10.1007/s00224-017-9770-0.
- 28 Dominik D. Freydenberger and Markus L. Schmid. Deterministic regular expressions with back-references. *J. Comput. Syst. Sci.*, 105:1–39, 2019. doi:10.1016/j.jcss.2019.04.001.
- 29 Jeffrey E. F. Friedl. *Mastering Regular Expressions*. O’Reilly, Sebastopol, CA, third edition, 2006.

- 30 Pawel Gawrychowski, Florin Manea, and Stefan Siemer. Matching patterns with variables under hamming distance. *CoRR*, abs/2106.06249, 2021. [arXiv:2106.06249](#).
- 31 Pawel Gawrychowski and Przemyslaw Uznanski. Optimal trade-offs for pattern matching with k mismatches. *CoRR*, abs/1704.01311, 2017. [arXiv:1704.01311](#).
- 32 Pawel Gawrychowski and Przemyslaw Uznanski. Towards unified approximate pattern matching for hamming and l_1 distance. In *Proc. 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018*, volume 107 of *LIPICs*, pages 62:1–62:13, 2018. doi:10.4230/LIPICs.ICALP.2018.62.
- 33 Juha Kärkkäinen and Peter Sanders. Simple linear work suffix array construction. In *Proc. 30th International Colloquium Automata, Languages and Programming, ICALP 2003*, volume 2719 of *Lecture Notes in Computer Science*, pages 943–955, 2003. doi:10.1007/3-540-45061-0_73.
- 34 Juha Kärkkäinen, Peter Sanders, and Stefan Burkhardt. Linear work suffix array construction. *J. ACM*, 53(6):918–936, 2006. doi:10.1145/1217856.1217858.
- 35 Gad M. Landau and Uzi Vishkin. Efficient string matching in the presence of errors. In *Proc. 26th Annual Symposium on Foundations of Computer Science, FOCS 1985*, pages 126–136, 1985. doi:10.1109/SFCS.1985.22.
- 36 Ming Li, Bin Ma, and Lusheng Wang. Finding similar regions in many sequences. *J. Comput. Syst. Sci.*, 65(1):73–96, 2002. doi:10.1006/jcss.2002.1823.
- 37 M. Lothaire. *Combinatorics on Words*. Cambridge University Press, 1997. doi:10.1017/CB09780511566097.
- 38 M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002. doi:10.1017/CB09781107326019.
- 39 Florin Manea and Markus L. Schmid. Matching patterns with variables. In *Proc. 12th International Conference Combinatorics on Words, WORDS 2019*, volume 11682 of *Lecture Notes in Computer Science*, pages 1–27, 2019. doi:10.1007/978-3-030-28796-2_1.
- 40 Dániel Marx. Closest substring problems with small distances. *SIAM J. Comput.*, 38(4):1382–1410, 2008. doi:10.1137/060673898.
- 41 Eugene W. Myers and Webb Miller. Approximate matching of regular expressions. *Bull. Math. Biol.*, 51(1):5–37, 1989. doi:10.1007/BF02458834.
- 42 Daniel Reidenbach and Markus L. Schmid. Patterns with bounded treewidth. *Inf. Comput.*, 239:87–99, 2014. doi:10.1016/j.ic.2014.08.010.
- 43 Markus L. Schmid. A note on the complexity of matching patterns with variables. *Inf. Process. Lett.*, 113(19):729–733, 2013. doi:10.1016/j.ipl.2013.06.011.
- 44 Markus L. Schmid and Nicole Schweikardt. A purely regular approach to non-regular core spanners. In *Proc. 24th International Conference on Database Theory, ICDT 2021*, volume 186 of *LIPICs*, pages 4:1–4:19, 2021. doi:10.4230/LIPICs.ICDT.2021.4.
- 45 Takeshi Shinohara. Polynomial time inference of pattern languages and its application. In *Proc. 7th IBM Symposium on Mathematical Foundations of Computer Science, MFCS*, pages 191–209, 1982.
- 46 Takeshi Shinohara and Setsuo Arikawa. Pattern inference. In *Algorithmic Learning for Knowledge-Based Systems, GOSLER Final Report*, volume 961 of *LNAI*, pages 259–291, 1995.
- 47 Przemyslaw Uznanski. Recent advances in text-to-pattern distance algorithms. In *Proc. 16th Conference on Computability in Europe, CiE 2020*, volume 12098 of *Lecture Notes in Computer Science*, pages 353–365, 2020. doi:10.1007/978-3-030-51466-2_32.
- 48 Ryan Williams. A new algorithm for optimal 2-constraint satisfaction and its implications. *Theor. Comput. Sci.*, 348(2-3):357–365, 2005. doi:10.1016/j.tcs.2005.09.023.

A Computational Model

The computational model we use to describe our results is the standard unit-cost RAM with logarithmic word size: for an input of size n , each memory word can hold $\log n$ bits. Arithmetic and bitwise operations with numbers in $[1 : n]$ are, thus, assumed to take $O(1)$

time. Numbers larger than n , with ℓ bits, are represented in $O(\ell/\log n)$ memory words, and working with them takes time proportional to the number of memory words on which they are represented. In all the problems, we assume that we are given a word w and a pattern α , with $|w| = n$ and $|\alpha| = m \leq n$, over a terminal-alphabet $\Sigma = \{1, 2, \dots, \sigma\}$, with $|\Sigma| = \sigma \leq n$. The variables are chosen from the set $\{x_1, \dots, x_n\}$ and can be encoded as integers between $n + 1$ and $2n$. That is, we assume that the processed words are sequences of integers (called letters or symbols), each fitting in $O(1)$ memory words. This is a common assumption in string algorithms: the input alphabet is said to be *an integer alphabet*. For instance, the same assumption was also used for developing efficient algorithms for `Match` in [21]. For a more detailed general discussion on this model see, e.g., [17].

B Proofs

► **Lemma 3.1.** *Let w and u , with $|w| = |u| = n$, be two words and δ a non-negative integer. Assume that, in a preprocessing phase, we have constructed $LCS_{w,u}$ -data structures. We can compute $\min(\delta + 1, d_{\text{HAM}}(u, w))$ using $\delta + 1$ $LCS_{w,u}$ queries, so in $O(\delta)$ time.*

Proof. Let $a = b = m$ and $d = 0$. While $a > 0$ and $d \leq \delta$ execute the following steps. Compute $h = LCS_{w,u}(a, b)$. If $h < b$, then increment d by 1, set $a \leftarrow a - h - 1$ and $b \leftarrow b - h - 1$, and start another iteration of the while-loop. If $h = b$, then set $b \leftarrow 0$ and exit the while-loop.

It is not hard to note that before each iteration of the while loop it holds that $d = d_{\text{HAM}}(w[a + 1 : m], u[b + 1 : m])$. When the while loop is finished, $d = \min(d_{\text{HAM}}(w[i - m + 1 : i], u[1 : m]), \delta + 1)$. In each iteration we first identify the length h of the longest common suffix of $w[1 : a]$ and $u[1 : b]$. Then, we jump over this suffix, as it causes no mismatches, and have either traversed completely the words w and u (and we do not need to do anything more), or we have reached a mismatch between w and u , on position $a - h = b - h$. In the latter case, we count this mismatch, jump over it, and repeat the process (but only if the number of mismatches is still at most δ). So, in other words, we go through the mismatches of w and u , from right to left, and jump from one to the next one using $LCS_{w,u}$ queries. If we have more than δ mismatches, we do not count all of them, but stop as soon as we have met the $(\delta + 1)^{\text{th}}$ mismatch. Accordingly, the algorithm is correct. Clearly, we only need $\delta + 1$ $LCS_{w,u}$ -queries and the time complexity of this algorithm is $O(\delta)$, once the $LCS_{w,u}$ -data structures are constructed. ◀

► **Lemma 3.2.** *Given a word w , with $|w| = n$, a word u , with $|u| = m < n$, and a non-negative integer δ , we can compute in $O(n\delta)$ time the array $D[m : n]$ with $n - m + 1$ elements, where $D[i] = \min(\delta + 1, d_{\text{HAM}}(w[i - m + 1 : i], u))$.*

Proof. We first construct, in linear time, the $LCS_{w,u}$ -data structures for the input words. Note that the $LCS_{w,u}$ -data structure can be directly used as $LCS_{w[i:i+m-1],u}$ data structure, for all $i \leq n - m + 1$.

Then, for each position i of w , with $i \leq m$, we use Lemma 3.1 to compute, in $O(\delta)$ time the value $d = \min(d_{\text{HAM}}(u, w[i - m + 1 : i]), \delta + 1)$. We then set $D[i] \leftarrow d$. By the correctness of Lemma 3.1, we get the correctness of this algorithm. Clearly, its time complexity is $O(n\delta)$. ◀

► **Theorem 4.4.** *$\text{MinMisMatch}_{1\text{RepVar}}$ and $\text{MisMatch}_{1\text{RepVar}}$ can be solved in $O(n^{k+2}m)$ time, where k is the number of x -blocks in the input pattern α .*

Proof. Once more, we only show how $\text{MinMismatch}_{1\text{RepVar}}$ can be solved. The result for $\text{Mismatch}_{1\text{RepVar}}$ follows then immediately.

In $\text{MinMismatch}_{1\text{RepVar}}$, we are given a word w , of length n , and a pattern α , of length m , which, as stated above, has exactly k x -blocks. Thus $\alpha = \prod_{i=1}^k (\gamma_{i-1} \beta_i) \gamma_k$, where the factors β_i , for $i \in \{1, \dots, k\}$, are the x -blocks of α . It is easy to observe that $\text{var}(\gamma_i) \cap \text{var}(\gamma_j) = \emptyset$, for all i and j , and $\gamma = \gamma_0 \gamma_1 \cdots \gamma_k$ is a regular pattern.

When aligning α to w we actually align each of the patterns γ_j and β_i , for $0 \leq j \leq k$ and $1 \leq i \leq k$, to respective factors of the word w . Moreover, the factors to which these patterns are respectively aligned are completely determined by the length ℓ of the image of x , and the starting positions h_i of the factors aligned to the patterns β_i , for $1 \leq i \leq k$. Knowing the length ℓ of the image of x , we can also compute, for $1 \leq i \leq k$, the length ℓ_i of β_i , when x is replaced by a string of length ℓ . In this case, γ_0 is aligned $u_0 = w[1..h_1 - 1]$ and, for $1 \leq i \leq k$, β_i is aligned to $w_i = w[h_i : h_i + \ell_i - 1]$ and γ_i is aligned $u_i = w[h_i + \ell_i : h_{i+1} - 1]$ (where $h_{k+1} = n + 1$). Thus, $\beta_1 \cdots \beta_k$ matches $w_1 \cdots w_k$ and we can use Theorem 4.2 to determine $d_{\text{HAM}}(\beta_1 \cdots \beta_k, w_1 \cdots w_k)$ (or, in other words, determine the string u_x that should replace x in order to realize this Hamming distance). Further, we can use Theorem 3.4 to compute $d_{\text{HAM}}(\gamma_i, u_i)$, for all $i \in \{0, \dots, k\}$. Adding all these distances up, we obtain a total distance $D_{\ell, h_1, \dots, h_k}$; this value depends on ℓ, h_1, \dots, h_k .

So, we can simply iterate over all possible choices for ℓ, h_1, \dots, h_k and find $d_{\text{HAM}}(\alpha, w)$ as the minimum of the numbers $D_{\ell, h_1, \dots, h_k}$.

By the explanations above, it is straightforward that the approach is correct: we simply try all possibilities of aligning α with w . The time complexity is, for each choice of ℓ, h_1, \dots, h_k , $O(\sum_{i=1}^k |w_i|) \subseteq O(n)$ for the part corresponding to the computation of the optimal alignment between the factors β_i and the words w_i , and $O(\sum_{i=0}^k |u_i| d_{\text{HAM}}(\gamma_i, u_i)) \subseteq O(nm)$ for the part corresponding to the computation of the optimal alignment between the factors γ_i and the words u_i . So, the overall complexity of this algorithm is $O(n^{k+2}m)$. ◀

► **Theorem 4.5.** $\text{MinMismatch}_{\text{kLOC}}$ and $\text{Mismatch}_{\text{kLOC}}$ can be solved in $O(n^{2k+2}m)$ time.

Proof. We only present the solution for $\text{MinMismatch}_{\text{kLOC}}$ (as it trivially works in the case of $\text{Mismatch}_{\text{kLOC}}$ too).

Let us note that, by the results in [18], we can compute a marking sequence of α in $O(m^{2k}k)$ time. So, after such a preprocessing phase, we can assume that we have a word w , a k -local pattern α (with p variables) with a witness marking sequence $x_1 \leq \dots \leq x_p$ for the k -locality of α , and we want to compute $d_{\text{HAM}}(\alpha, w)$.

Generally, the main idea behind matching kLOC -patterns is that when looking for possible ways to align such a pattern α to a word w we can consider the variables in the order given by the marking sequence, and, when reaching variable x_i , we try all possible assignments for x_i . The critical observation here is that after each such assignment of a new variable, we only need to keep track of the way the $t \leq k$ length-maximal factors of α , which contain only marked variables and terminals, match (at most) $t \leq k$ factors of w .

We will use this approach in our algorithm for $\text{MinMismatch}_{\text{kLOC}}$.

The first step of this algorithm is the following. We go through α and identify all x_1 -blocks: $\beta_{1,1}, \dots, \beta_{1,j_1}$. Because α is k -local, we have that $j_1 \leq k$. For each $2j_1$ -tuple (i_1, \dots, i_{2j_1}) of positions of w , we compute the minimum number of mismatches if we align (simultaneously) the patterns β_g to the factors $w[i_{2g-1} : i_{2g}]$, for g from 1 to j_1 , respectively. This reduces to finding an assignment for x_1 which aligns optimally the patterns $\beta_{1,g}$ to the respective factors, and can be done in $O(n)$ time using Theorem 4.2. For each $2j_1$ -tuple (i_1, \dots, i_{2j_1}) of

positions of w , we denote by $M_1(i_1, \dots, i_{2j_1})$ the minimum number of mismatches resulting from the (simultaneous) alignment of the patterns $\beta_{1,g}$ to the factors $w[i_{2g-1} : i_{2g}]$, for g from 1 to j_1 , respectively. Clearly, M_1 can be seen as a j_1 -dimensional array.

Assume that after $h \geq 1$ steps of our algorithm we have computed the factors $\beta_{h,1}, \dots, \beta_{h,j_h}$ of α , which are length-maximal factors of α which only contain the variables x_1, \dots, x_h and terminals (i.e., extending them to the left or right would introduce a new variable x_ℓ with $\ell > h$); as α is k -local, we have $j_h \leq k$. Moreover, for each $2j_h$ -tuple (i_1, \dots, i_{2j_h}) of positions of w , we have computed $M_h(i_1, \dots, i_{2j_h})$, the minimum number of mismatches if we align (simultaneously) the patterns $\beta_{h,g}$ to the factors $w[i_{2g-1} : i_{2g}]$, for g from 1 to j_h , respectively. M_h is implemented as a j_h dimensional array, and this assumption clearly holds after the first step.

We now explain how step $h + 1$ is performed.

1. We compute the factors $\beta_{h+1,1}, \dots, \beta_{h+1,j_{h+1}}$ of α , which are length-maximal factors of α which only contain the variables x_1, \dots, x_{h+1} and terminals (i.e., extending them to the left or right would introduce a new variable x_ℓ with $\ell > h + 1$). Clearly, $\beta_{h+1,r}$ is either an x_{h+1} -block or it has the form $\beta_{h+1,r} = \gamma_{r,0}\beta_{h,a_r}\gamma_{r,1} \cdots \beta_{r,a_r+b_r}\gamma_{r,b_r+1}$ where the patterns $\gamma_{r,t}$ contain only the variable x_{h+1} and terminals and extending $\beta_{h+1,r}$ to the left or right would introduce a new variable x_ℓ with $\ell > h + 1$.
2. We initialize the values $M_{h+1}(i_1, \dots, i_{2j_{h+1}}) \leftarrow \infty$, for each $2j_{h+1}$ -tuple $(i_1, \dots, i_{2j_{h+1}})$ of positions of w .
3. For each $\ell \leq n$ (where ℓ corresponds to the length of the image of x_{h+1}) and each $2j_h$ -tuple (i_1, \dots, i_{2j_h}) of positions of w such that $M_h(i_1, \dots, i_{2j_h})$ is finite do the following:
 - a. We compute the tuple $(i'_1, \dots, i'_{2j_{h+1}})$ such that $\beta_{h+1,g}$ is aligned to the factor $w[i'_{2g-1} : i'_{2g}]$, for g from 1 to j_{h+1} , respectively. This can be computed based on the fact that the factors $\beta_{h,g}$ are aligned to the factors $w[i_{2g-1} : i_{2g}]$, for g from 1 to j_h , respectively, and the image of x_{h+1} has length ℓ .
 - b. We compute the factors of w aligned to x_{h+1} in the alignment computed in the previous line. Then, we can use the algorithm from Theorem 4.2 and the value of $M_h(i_1, \dots, i_{2j_h})$ to compute an assignment for x_{h+1} which aligns optimally the patterns $\beta_{h+1,g}$ to the corresponding factors of w .
 - c. If the number of the mismatches in this alignment is smaller than the current value of $M_{h+1}(i'_1, \dots, i'_{2j_{h+1}})$, we update $M_{h+1}(i'_1, \dots, i'_{2j_{h+1}})$.

This dynamic programming approach is clearly correct. In $M_{h+1}(i_1, \dots, i_{2j_{h+1}})$ we have the optimal alignment of the patterns $\beta_{h+1,1}, \dots, \beta_{h+1,j_{h+1}}$ to $w[i_1 : i_2], \dots, w[i_{2j_{h+1}-1} : i_{2j_{h+1}}]$, respectively. As far as the complexity is concerned, the lines 1, 3.a, 3.b, 3.c can be implemented in linear time, while the for-loop is iterated $O(n^{2k+1})$ times. Line 2 takes $O(n^{2k})$ times. The whole computation in step $h + 1$ of the algorithm takes, thus, $O(n^{2k+1})$ time.

Now, we execute the procedure described above for h from 2 to m , and, in the end, we compute the array M_m . The answer to our instance of the problem $\text{MinMismatch}_{\text{KLOC}}$ is $M_m(1, n)$. The overall time complexity needed to perform this computation is $O(mn^{2k+1})$ time. The preprocessing phase, in which the marking sequence and the array M_1 were computed, takes also $O(mn^{2k+1})$ time. So, the complexity stated in the statement is reached by our algorithm. \blacktriangleleft

► Theorem 4.6. *For each constant $r \geq 3$, there exists an algorithm with run-time $O(n^{r+3})$ for $\text{MinMismatch}_{1\text{RepVar}}$ whose output distance is at most $\min \left\{ 2, \left(1 + \frac{4\sigma-4}{\sqrt{e}(\sqrt{4r+1}-3)} \right) \right\} d_{\text{HAM}}(\alpha, w)$.*

Proof. We first note that there exists a relatively simple algorithm solving $\text{MinMisMatch}_{1\text{RepVar}}$ such that the output distance is no more than $2d_{\text{HAM}}(\alpha, w)$ (which also works for integer alphabets).

Indeed, assume that we have a substitution h for which $d_{\text{HAM}}(h(\alpha), w) = d_{\text{HAM}}(\alpha, w)$. Assume that the repeated variable x is mapped by h to a string u and the t occurrences of x are aligned, under h , to the factors w_1, w_2, \dots, w_t of w . Now, let w_i be such $d_{\text{HAM}}(u, w_i) \leq d_{\text{HAM}}(u, w_j)$ for all $j \neq i$. Let us consider now the substitution h' which substitutes x by w_i and all the other variables exactly as h did. We claim that $d_{\text{HAM}}(h'(\alpha), u) \leq 2d_{\text{HAM}}(h(\alpha), u)$. It is easy to see that $d_{\text{HAM}}(h'(\alpha), w) - d_{\text{HAM}}(h(\alpha), w) = \sum_{j=i}^t (d_{\text{HAM}}(w_i, w_j) - d_{\text{HAM}}(u, w_j)) \leq \sum_{j=i}^t (d_{\text{HAM}}(w_i, u) + d_{\text{HAM}}(u, w_j) - d_{\text{HAM}}(u, w_i))$ (where the last inequality follows from the triangle inequality for the Hamming distance). Thus, $d_{\text{HAM}}(h'(\alpha), w) - d_{\text{HAM}}(h(\alpha), w) \leq \sum_{j=i}^t d_{\text{HAM}}(w_i, u) \leq \sum_{j=i}^t d_{\text{HAM}}(w_j, u) \leq d_{\text{HAM}}(h(\alpha), u)$. So our claim holds.

A consequence of the previous observation is that there exists a substitution h' that maps x to a factor of w and produces a string $h'(\alpha)$ such that $d_{\text{HAM}}(h'(\alpha), u) \leq 2d_{\text{HAM}}(\alpha, u)$. So, for each factor u of w , we x by u in α to obtain a regular pattern α' , then use Theorem 3.4 to compute $d_{\text{HAM}}(\alpha', w)$. We return the smallest value $d_{\text{HAM}}(\alpha', w)$ achieved in this way. Clearly, this is at most $2d_{\text{HAM}}(\alpha, w)$. The complexity of this algorithm is $O(n^4)$, as it simply uses the quadratic algorithm of Theorem 3.4 for each factor of w .

We will now show how this algorithm can be modified to produce a value closer to $d_{\text{HAM}}(\alpha, w)$, while being less efficient.

The algorithm consists of the following main steps:

1. For $\ell \leq n/r$ and r factors u_1, \dots, u_r of length ℓ of w do the following:
 - a. Compute u_{u_1, \dots, u_r} the median string of u_1, \dots, u_r using Lemma 4.1.
 - b. Let α' be the regular pattern obtained by replacing x by u_{u_1, \dots, u_r} in α .
 - c. Compute the distance $d_{u_1, \dots, u_r} = d_{\text{HAM}}(\alpha', w)$ using Theorem 3.4.
2. Return the smallest distance d_{u_1, \dots, u_r} computed in the loop above.

Clearly, for $r = 1$ the above algorithm corresponds to the simple algorithm presented in the beginning of this proof. Let us analyse its performance for an arbitrary choice of r .

The complexity is easy to compute: we need to consider all possible choices for ℓ and the starting positions of u_1, \dots, u_r . So, we have $O(n^{r+1})$ possibilities to select the non-overlapping factors u_1, \dots, u_r of length ℓ of w . The computation done inside the loop can be performed in $O(n^2)$ time. So, overall, our algorithm runs in $O(n^{r+3})$ time.

Now, we want to estimate how far away from $d_{\text{HAM}}(\alpha, w)$ is the value this algorithm returns. In this case, we will make use of the fact that the input terminal-alphabet is constant. We follow closely (and adapt to our setting) the approach from [36].

Firstly, a notation. In step 1.b of the algorithm above, we align α' to w with a minimal number of mismatches. In this alignment, let d'_{u_1, \dots, u_r} be the total number of mismatches caused by the factors u_{u_1, \dots, u_r} which replaced the occurrences of the variable x in α .

Now, assume that we have a substitution h for which $d_{\text{HAM}}(h(\alpha), w) = d_{\text{HAM}}(\alpha, w) = d_{\text{opt}}$. Assume also that the repeated variable x is mapped by h to a string u_{opt} of length L and the t occurrences of x are aligned, under h , to the factors w_1, w_2, \dots, w_t of w . Let d'_{opt} be the number of mismatches caused by the alignment of the images of the t occurrences of x under h to the factors w_1, w_2, \dots, w_t . Finally, let $\rho = 1 + \frac{4\sigma-4}{\sqrt{e}(\sqrt{4r+1}-3)}$.

Note that, for $\ell = L$, u_1, \dots, u_r correspond to a set of randomly chosen numbers i_1, \dots, i_r from $\{1, \dots, n\}$: their starting positions. We will show in the following that $E[d'_{u_1, \dots, u_r}] \leq \rho d'_{\text{opt}}$. If this inequality holds, then we can apply the probabilistic method: there exists at least a choice of u_1, \dots, u_r of length L such that $d'_{u_1, \dots, u_r} \leq \rho d'_{\text{opt}}$. As we try all possible lengths ℓ and all variants for choosing u_1, \dots, u_r of length ℓ , we will also consider

the choice of u_1, \dots, u_r of length L such that $d'_{u_1, \dots, u_r} \leq \rho d'_{opt}$, and it is immediate that, for that, for the respective u_1, \dots, u_r we also have that $d_{u_1, \dots, u_r} \leq \rho d_{opt}$. Thus, the value returned by our algorithm is at most ρd_{opt} .

So, let us show the inequality $E[d'_{u_1, \dots, u_r}] \leq \rho d'_{opt}$.

For $\mathbf{a} \in \Sigma$, let $f_j(\mathbf{a}) = |\{i \mid 1 \leq i \leq t, w_i[j] = \mathbf{a}\}|$. Now, for an arbitrary string s of length L , we have that $\sum_{i=1}^t d_{\text{HAM}}(w_i, s) = \sum_{j=1}^L (t - f_j(s[j]))$. So, for $s = u_{opt}$ we get $\sum_{i=1}^t d_{\text{HAM}}(w_i, u_{opt}) = \sum_{j=1}^L (t - f_j(u_{opt}[j]))$, and for $s = u_{u_1, \dots, u_r}$ we have that $d'_{opt} = \sum_{i=1}^t d_{\text{HAM}}(w_i, u_{u_1, \dots, u_r}) = \sum_{j=1}^L (t - f_j(u_{u_1, \dots, u_r}[j]))$.

Therefore, $E[d'_{u_1, \dots, u_r}] = E\left[\sum_{j=1}^L (t - f_j(u_{u_1, \dots, u_r}[j]))\right] = \sum_{j=1}^L E[t - f_j(u_{u_1, \dots, u_r}[j])]$.

Consequently, $E[d'_{u_1, \dots, u_r} - d'_{opt}] = \sum_{j=1}^L (E[t - f_j(u_{u_1, \dots, u_r}[j])] - t + f_j(u_{opt}[j]))$.

That is, $E[d'_{u_1, \dots, u_r} - d'_{opt}] = \sum_{j=1}^L E[f_j(u_{opt}[j]) - f_j(u_{u_1, \dots, u_r}[j])]$.

By Lemma 7 of [36], we have that $E[f_j(u_{opt}[j]) - f_j(u_{u_1, \dots, u_r}[j])] \leq (\rho - 1)(t - f_j(u_{opt}[j]))$.

Hence, $E[d'_{u_1, \dots, u_r} - d'_{opt}] \leq (\rho - 1) \sum_{j=1}^L (t - f_j(u_{opt}[j])) = (\rho - 1)d'_{opt}$.

So, we indeed have that $E[d'_{u_1, \dots, u_r}] \leq \rho d'_{opt}$.

In conclusion, the statement of the theorem holds. \blacktriangleleft

► **Theorem 4.8.** $\text{MinMisMatch}_{1\text{RepVar}}$ has no EPTAS unless $FPT = W[1]$.

Proof. Assume, for the sake of a contradiction, that $\text{MinMisMatch}_{1\text{RepVar}}$ has an EPTAS. That is, for an input word w and an 1RepVar -pattern α , there exists a polynomial time algorithm which returns as answer to $\text{MinMisMatch}_{1\text{RepVar}}$ a value $\delta' \leq (1 + \epsilon)d_{\text{HAM}}(\alpha, w)$, and the exponent of the polynomial in its running time is independent of ϵ .

An algorithm for minCP would first implement the reduction in Theorem 4.7 to obtain a word w and a pattern α . Then it uses the EPTAS for $\text{MinMisMatch}_{1\text{RepVar}}$ to approximate the distance between α and w with approximation ratio $(1 + \frac{\epsilon}{2m})$. Assuming that this EPTAS returns the value D , the answer returned by this algorithm for the minCP problem is $D - m$.

As explained in the proof of Theorem 4.7, it is easy to see that the distance between the word w and the pattern α constructed in the respective reduction is $m + \Delta$, if Δ is the answer to the instance of the minCP problem. Thus, the value D returned by the EPTAS for $\text{MinMisMatch}_{1\text{RepVar}}$ fulfils $m + \Delta \leq D \leq (1 + \frac{\epsilon}{2m})(m + \Delta)$. So, we have $\Delta \leq D - m \leq \frac{\epsilon}{2} + (1 + \frac{\epsilon}{2m})\Delta$. We get that $\Delta \leq D - m \leq (1 + \frac{\epsilon}{2m} + \frac{\epsilon}{2\Delta})\Delta \leq (1 + \epsilon)\Delta$. So, indeed, $D - m$ would be a $(1 + \epsilon)$ -approximation of Δ .

Therefore, this would yield an EPTAS for minCP . This is a contradiction to the results reported in [7], where it was shown that such an EPTAS does not exist, unless $FPT = W[1]$. This concludes our proof. \blacktriangleleft