

The Smoothed Complexity of Edit Distance*

Alexandr Andoni[†]
MIT
andoni@mit.edu

Robert Krauthgamer[‡]
The Weizmann Institute of Science
robert.krauthgamer@weizmann.ac.il

March 18, 2009

Abstract

We initiate the study of the smoothed complexity of sequence alignment, by proposing a semi-random model of edit distance between two input strings, generated as follows. First, an adversary chooses two binary strings of length d and a longest common subsequence A of them. Then, every character is perturbed independently with probability p , except that A is perturbed in exactly the same way inside the two strings.

We design two efficient algorithms that compute the edit distance on smoothed instances up to a constant factor approximation. The first algorithm runs in near-linear time, namely $d^{1+\varepsilon}$ for any fixed $\varepsilon > 0$. The second one runs in time sublinear in d , assuming the edit distance is not too small. These approximation and runtime guarantees are significantly better than the bounds known for worst-case inputs, e.g. near-linear time algorithm achieving approximation roughly $d^{1/3}$, due to Batu, Ergün, and Sahinalp [SODA 2006]¹.

Our technical contribution is twofold. First, we rely on finding matches between substrings in the two strings, where two substrings are considered a match if their edit distance is relatively small, a prevailing technique in commonly used heuristics, such as PatternHunter of Ma, Tromp and Li [Bioinformatics, 2002]. Second, we effectively reduce the smoothed edit distance to a simpler variant of (worst-case) edit distance, namely, edit distance on permutations (a.k.a. Ulam's metric). We are thus able to build on algorithms developed for the Ulam metric, whose much better algorithmic guarantees usually do not carry over to general edit distance.

1 Introduction

The *edit distance* (aka *Levenshtein distance*) between two strings is the number of insertions, deletions, and substitutions needed to transform one string into the other. This distance is of key importance in several fields, such as computational biology and text processing, and consequently computational problems involving the edit distance were studied extensively, both theoretically and experimentally, see e.g. the detailed survey on edit distance by Navarro [Nav01]. Despite extensive research, the worst-case guarantees currently known for algorithms dealing with edit distance

*An extended abstract of this article appeared in ICALP(I), 2008.

[†]Part of this work done while at IBM Almaden Research Center.

[‡]Part of this work done while at IBM Almaden Research Center. This research was supported in part by a grant from the Fufeld Research Fund, and by the Israel Science Foundation grant #452/08.

¹We note that, very recently, after the conference version of this paper appeared, this approximation was improved to $2^{\tilde{O}(\sqrt{\log d})}$ by Andoni and Onak.

are quite poor, especially in comparison to the Hamming distance (which is just the number of substitutions to transform one string into the other).

The most basic problem is to compute the edit distance between two strings of length d over alphabet Σ . The worst-case running time known for this problem has not improved in three decades — the problem can be solved using dynamic programming in time $O(d^2)$ [WF74], and in time $O(d^2/\log^2 d)$ when the alphabet has constant size [MP80]. Unfortunately, such near-quadratic time is prohibitive when working on large datasets, which is common in areas such as computational biology. The gold standard is to achieve a linear-time algorithm, or even sublinear in several cases, which has triggered the study of very efficient *distance estimation* algorithms – algorithms that compute an approximation to the edit distance. In particular, the best quasi-linear time algorithm, due to Batu, Ergün, and Sahinalp [BES06], achieves $d^{1/3+o(1)}$ approximation ² (improving over [BJKK04]), and the only known sublinear time algorithm, due to Batu, Ergün, Kilian, Magen, Raskhodnikova, Rubinfeld and Sami [BEK⁺03], decides whether the edit distance is $O(d^\alpha)$ or $\Omega(d)$ in time $O(d^{\max\{\alpha/2, 1-2\alpha\}})$. In fact, distance estimation with sublogarithmic approximation factor was recently proved impossible in a certain model of low communication complexity [AK07]. In practice, this situation is mitigated by heuristic algorithms. In computational biology settings for instance, tools such as BLAST [AGM⁺90] are commonly used to solve the problem quickly, essentially by relying on heuristic considerations that sacrifice some sensitivity.

We initiate the study of the smoothed complexity of sequence alignment, by proposing a semi-random model of edit distance (the input is a worst-case instance modified by a random perturbation), and design for it very efficient approximation algorithms. Specifically, an adversary chooses two strings and a longest common subsequence of them, and every character is perturbed independently with probability $0 \leq p \leq 1$, except that every character in the common subsequence is perturbed in the same way in the two strings. Semi-random models appeared in the literature in other contexts, but to the best of our knowledge, not for sequence alignment problems; see Section 1.2 for more details. Our algorithms for the smoothed model approximate the edit distance within a constant factor in linear, and even sublinear time.

Why study semi-random models of sequence alignment? First, they elude the extreme difficulty posed by worst-case inputs, while avoiding the naivete of average-case (random) inputs. Using these models as a theoretical testbed for practical algorithms may lead to designing new algorithmic techniques, and/or to providing rigorous explanation for the empirical success of well-known heuristics. Second, studying algorithms for semi-random models may be viewed as an attack on the worst-case complexity. It is difficult to quantify the progress we manage to make in this direction, but we certainly achieve much better performance guarantees on a very large collection of inputs (including random inputs as an extreme case), by delineating rather general assumptions on the input, under which we have efficient algorithms.

1.1 Our Contribution

A smoothed model. Let $0 < p \leq 1$ be a *perturbation probability*. In our smoothed model for edit distance, an input consisting of two strings, x and y , is generated as follows. (A more formal description is given in Section 1.3.)

1. An adversary chooses two strings $x^*, y^* \in \{0, 1\}^d$, and a longest common subsequence \mathcal{A} of

²We note that, very recently, after the conference version of this paper appeared, this approximation factor was improved to $2^{\tilde{O}(\sqrt{\log d})}$ by Andoni and Onak [AO09].

x^*, y^* .

2. Every character in x^* and y^* is replaced independently with probability p by a random bit, except that the perturbation of \mathcal{A} inside x and that of \mathcal{A} inside y are identical.

Results. We start by investigating the typical properties of a smoothed instance (x, y) , including proving that the expected edit distance $\text{ed}(x, y)$ is comparable to that of the generating strings, $\text{ed}(x^*, y^*)$.

Our first result is a deterministic algorithm that approximates the edit distance within a constant factor, and its smoothed runtime complexity is near-linear. Specifically, for any desired $0 < \varepsilon < 1$, the algorithm always obtains $O(\frac{1}{\varepsilon p} \log \frac{1}{\varepsilon p})$ approximation, and with high probability over the randomness in the smoothing, runs in time $O(d^{1+\varepsilon})$. For comparison, the algorithm of Batu, Ergün, and Sahinalp [BES06] for worst-case inputs requires a similar running time of $O(d^{1+\varepsilon})$ and achieves approximation $d^{(1-\varepsilon)/3+o(1)}$.

Our second result is a sublinear time algorithm for smoothed instances. Specifically, for every desired $0 < \varepsilon < 1$, the algorithm computes a $O(\frac{1}{\varepsilon p} \log \frac{1}{\varepsilon p})$ approximation to $\text{ed}(x, y)$ in time $O(d^{1+\varepsilon}/\sqrt{\text{ed}(x, y)})$. For comparison, recall that the algorithm of Batu et al. [BEK⁺03] for worst-case inputs can only distinguish a polynomially large gap in the edit distance, and only at the highest regime $\Omega(d)$.

Techniques. Our algorithms are based on two new technical ideas. The first one is to find *matches* of blocks (substrings) of length $L = O(\frac{1}{p} \log d)$ between the two strings, where two blocks are considered a match if they are at a small edit distance (say εL). This same idea, but in a more heuristic form, is used by practical tools. In particular, PatternHunter [MTL02] uses such a notion of matches (to identify “seeds”), significantly improving over BLAST [AGM⁺90], which considers only identical blocks to be a match. Thus, our smoothed analysis may be viewed as giving some rigorous explanation for the empirical success of such techniques.

The second idea is to reduce the problem to edit distance on permutations (in worst-case), called in the literature *Ulam’s distance*, or the *Ulam metric*. Here and throughout, a *permutation* is a string in which every symbol appears at most once.³ The Ulam metric is a submetric of edit distance, but the algorithmic bounds known for it are significantly better than those for the general edit distance. In particular, Ulam’s distance between permutations of length d can be computed in linear time $O(d \log d)$, e.g. using Patience Sorting. The main challenge we overcome is to design a reduction that distorts distances by at most a constant factor. Indeed, there is an easy reduction with distortion $L = O(\frac{1}{p} \log d)$, that follows simply because with high probability, in each string, the blocks of length L are all distinct, see [CK06, Section 3.1].

1.2 Related Work

Average-case analysis of edit distance. Random models for edit distance were studied in two contexts, for pattern matching and for nearest neighbor searching. In the former, the text is typically assumed to be random, i.e., each character is chosen uniformly and independently from the alphabet, and the pattern is usually not assumed to be random. We refer the reader to the

³It is sometimes convenient, though not crucial, to use an alphabet Σ with size larger than d . We then define a permutation as a string whose characters are all distinct.

survey [Nav01, Section 5.3] for details and references. For nearest neighbor search, the average-case model is quite similar, see [NBYST01, GP06].

Our model is considerably more general than the random strings model. In particular, the average-case analysis often relies on the fact that no short substring of the text is identical to any substring of the pattern, to quickly “reject” most candidate matches. In fact, for distance estimation, it is easy to distinguish the case of two random strings from the case of two (worst-case) strings at a smaller edit distance — just choose one random block of logarithmic length in the first string and check whether it is close in edit distance to at least one block in the second string. We achieve a near-linear time algorithm for a more adversarial model, albeit by allowing constant factor approximation.

Smoothed complexity and semi-random models. Smoothed analysis was pioneered by Spielman and Teng [ST04] as a framework aimed to explain the practical success of heuristics that do not admit traditional worst-case analysis. They analyzed the simplex algorithm for linear programming, and since then researchers investigated the smoothed complexity of several other problems, mostly numerical ones, but also some discrete problems. An emerging principle in smoothed analysis is to perform *property-preserving* perturbation [ST03], example of which is our model. Specifically, our model may be seen as performing a perturbation of x^* and y^* that preserves the common subsequence \mathcal{A} .

In combinatorial optimization problems, smoothed analysis is closely related to an earlier notion of semi-random models, which were initiated by Blum and Spencer [BS95]. This research program encompasses several interesting questions, such as *what algorithmic techniques are most effective (spectral methods?)*, and *when is the optimum solution likely to be unique, hard to find, or easy to certify*, see e.g. [FM97, FK01] and the references therein.

To the best of our knowledge, smoothed analysis and/or semi-random models were not studied before for sequence alignment problems.

Distance estimation. Algorithms for distance estimation are studied also in other scenarios, using different notions of efficiency. One such model is the communication complexity model, where two parties are each given a string, and they wish to estimate the distance between their strings using low communication. The sketching model falls into this category, with further restriction to simultaneous communication protocols. A communication lower bound was recently proved in [AK07] for the edit distance metric, even on permutations, and it holds for approximations as large as $\Omega(\log d / \log \log d)$.

1.3 Preliminaries

Strings. Let x be a string of length d over alphabet Σ . A *position* in the string is an index $i \in [d]$, where throughout we let $[k] = \{1, 2, \dots, k\}$. We write $x[i]$ or x_i to denote the symbol appearing in position i in x . Let $[i : j]$ denote the sequence of positions $(i, i + 1, \dots, j)$. We write $x[i : j]$ or $x_{[i:j]}$ for the corresponding substring of x . A *block* is a substring, often of a predetermined length.

A variant of edit distance. Let x, y be two strings. Define $\underline{\text{ed}}(x, y)$ to be the minimum number of character insertions and deletions needed to transform x into y . Character substitution are not allowed, in contrast to $\text{ed}(x, y)$, but a substitution can be simulated by a deletion followed by an

insertion, and thus $\text{ed}(x, y) \leq \underline{\text{ed}}(x, y) \leq 2 \text{ed}(x, y)$. Observe that

$$\underline{\text{ed}}(x, y) = |x| + |y| - 2 \text{LCS}(x, y),$$

where $\text{LCS}(x, y)$ is the length of the longest common subsequence of x and y .

Alignments. For two strings x, y of length d , an *alignment* is a function $A : [d] \rightarrow [d] \cup \{\perp\}$ that is monotonically increasing on $A^{-1}([d])$ and satisfies $x[i] = y[A(i)]$ for all $i \in A^{-1}([d])$. Define the *length* (or *size*) of the alignment as $\text{len}(A) = |A^{-1}([d])|$, i.e., the number of positions in x that are matched by A . Let the *cost* of A be $\text{cost}(A) = 2(d - \text{len}(A)) = 2|A^{-1}(\perp)|$, i.e. the number of positions in x and in y that are not matched by A . Observe that an alignment between x and y corresponds exactly to a common subsequence to x and y . Thus, if A is an alignment between x and y , then

$$\text{cost}(A) = 2(d - \text{len}(A)) \geq 2d - 2 \text{LCS}(x, y) = \underline{\text{ed}}(x, y),$$

with equality if and only if A is an alignment of maximum length.

Block matches. Consider two strings x, y and a block length $L \in [d]$. For blocks $x_{[i:i+L-1]}$ and $y_{[j:j+L-1]}$ of length L , we let $\text{ed}_A(x_{[i:i+L-1]}, y_{[j:j+L-1]})$ be the number of positions $k \in [i : i + L - 1]$ such that $A(k) \notin [j : j + L - 1]$. We let $\text{match}(x_{[i:i+L-1]})$ denote the block $y_{[j:j+L-1]}$, where $j \in [d - L + 1]$ minimizes $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]})$, breaking the ties arbitrarily. For an alignment A between x and y , let $\text{match}_A(i, L)$ be the block $y_{[j:j+L-1]}$, where $j \in [d - L + 1]$ minimizes $\text{ed}_A(x_{[i:i+L-1]}, y_{[j:j+L-1]})$, breaking the ties arbitrarily. Slightly abusing notation, we sometimes let match and match_A represent the corresponding position j (instead of the substring $y_{[j:j+L-1]}$), but the distinction will be clear from the context.

Smoothed model. Let $0 \leq p \leq 1$, let $x^*, y^* \in \{0, 1\}^d$ be two strings, and fix a maximum-length alignment A^* between x^* and y^* . Let $x, y \in \{0, 1\}^d$ be the strings obtained from x^*, y^* respectively, by replacing, independently with probability p , each character with a random one, except that the positions aligned by A^* are kept correlated. Formally, let $\pi_x \in \{0, 1\}^d$ be a string where each $\pi_x[j]$ is drawn independently to be 1 with probability $p/2$ and 0 otherwise, and let π_y be defined similarly (and independently), except for position $j \in A^*([d])$, for which we set $\pi_y[j] = \pi_x[(A^*)^{-1}(j)]$. Now let $x[i] = x^*[i] + \pi_x[i]$ and $y[i] = y^*[i] + \pi_y[i]$, where addition is done modulo 2. We call the pair (x, y) a *smoothed instance* of edit distance, and denote its distribution by $\text{Smooth}_p(x^*, y^*, A^*)$.

2 Typical properties of smoothed instances

We first show that the edit distance of a smoothed instance is likely to be similar to that of the strings used to generate it. We then turn our attention to the distance between different substrings of the smoothed strings x and y . Specifically, we show that blocks of length $L = O(p^{-1} \log d)$ are likely to be far from each other in terms of edit distance, with the few obvious exceptions of overlapping blocks and blocks that are aligned via the original alignment A^* .

Besides the inherent interest, these bounds are useful in the smoothed analysis of our algorithms carried out in subsequent sections.

2.1 Edit Distance of a Smoothed Instance

Theorem 2.1. *Let A^* be an optimal alignment between $x^*, y^* \in \{0, 1\}^d$, and fix $0 < p \leq 1$. Then a smoothed instance $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$ satisfies*

$$\Pr_{(x,y)} \left[\Omega\left(\frac{p}{\log(2/p)}\right) \text{ed}(x^*, y^*) \leq \text{ed}(x, y) \leq \text{ed}(x^*, y^*) \right] \geq 1 - 2^{-\Omega(p) \text{ed}(x^*, y^*)}.$$

Proof. Observe that $\text{ed}(x, y) \leq \text{ed}(x^*, y^*)$ always holds (i.e. with probability 1). We proceed to show that with high probability, $\underline{\text{ed}}(x, y) \geq \Omega\left(\frac{p}{\log(2/p)}\right) \cdot \underline{\text{ed}}(x^*, y^*)$, which by the facts from Section 1.3 would complete the proof. We let U denote the unaligned positions in x under A^* , i.e. $U = (A^*)^{-1}(\perp)$ and $|U| = \frac{1}{2} \underline{\text{ed}}(x^*, y^*)$.

Consider a *potential alignment* A between x and y , i.e. a map $A : [d] \mapsto [d] \cup \{\perp\}$ that is monotonically increasing on $A^{-1}([d])$, and suppose that $\text{cost}(A) = 2|A^{-1}(\perp)|$ is at most $\alpha \cdot \underline{\text{ed}}(x^*, y^*)$ for a small $0 < \alpha \leq 1/4$ to be chosen later. For A to be an actual alignment, we must additionally have that $x[i] = y[A(i)]$ for every position $i \notin A^{-1}(\perp)$, and in particular for every position $i \in U \setminus A^{-1}(\perp)$. The number of such positions is at least $|U| - |A^{-1}(\perp)| \geq \frac{1}{2} \underline{\text{ed}}(x^*, y^*) - \frac{1}{2} \alpha \underline{\text{ed}}(x^*, y^*) \geq \frac{1}{4} \underline{\text{ed}}(x^*, y^*)$. For each of them, $x^*[i]$ is perturbed independently of $y^*[A(i)]$, and thus $x[i] \neq y[A(i)]$ occurs with probability at least $p/2$. These events might not be mutually independent due to correlations via A^* , but it is easy to see that for at least half of such i , the probability is at least $p/2$ even when conditioned on earlier events (namely, $x[i]$ is independent of $x[i']$ and $y[A(i')]$ for all $i' < i$). Thus, the probability that A is an actual alignment is at most

$$\Pr \left[x[i] = y[A(i)] \text{ for all } i \in U \setminus A^{-1}(\perp) \right] \leq \left(1 - \frac{p}{2}\right)^{\underline{\text{ed}}(x^*, y^*)/8} \leq e^{-p \underline{\text{ed}}(x^*, y^*)/16}.$$

We will apply a union bound on all potential alignments, and thus it suffices to have an upper bound on the number of different values taken by $A|_U$, the restriction of A to the positions in U . We note that $A|_U$ is determined by the number of insertions and deletions occurring between every two successive positions in U (including the insertions and deletions before the first position in U and after the last position in U). Thus we can count the number of $A|_U$ as:

$$\#\{A|_U\} \leq \binom{|U| + \frac{1}{2}\alpha \underline{\text{ed}}(x^*, y^*)}{\frac{1}{2}\alpha \underline{\text{ed}}(x^*, y^*)}^3 \leq \left(\frac{e(1+\alpha)}{\alpha}\right)^{1.5\alpha \underline{\text{ed}}(x^*, y^*)} \leq \left(\frac{1}{\alpha^2}\right)^{1.5\alpha \underline{\text{ed}}(x^*, y^*)}.$$

Choosing $\alpha = \frac{cp}{\log(2/p)}$ for a sufficiently small constant $c > 0$, we get by a union bound that

$$\Pr \left[\underline{\text{ed}}(x, y) \leq \alpha \underline{\text{ed}}(x^*, y^*) \right] \leq e^{[3\alpha \ln(1/\alpha) - p/16] \cdot \underline{\text{ed}}(x^*, y^*)} \leq e^{-(p/32) \underline{\text{ed}}(x^*, y^*)}$$

□

2.2 Edit Distance Between Different Blocks

Our next lemma gives properties of typical distances between two blocks from x and y . Although, in spirit, this lemma is similar to Theorem 2.1, it differs in that, in this lemma, we consider blocks whose perturbations are correlated, e.g., overlapping blocks in the same string. This technical difficulty impedes direct concentration bounds used in the previous theorem, and thus will require more ideas to complete the proof.

Lemma 2.2. Let A^* be an optimal alignment between $x^*, y^* \in \{0, 1\}^d$ and fix $0 < p \leq 1$. Let $L \geq \frac{C}{p} \log d$ for a sufficiently large constant $C > 0$, and let $c_a, c_b, c_c > 0$ be sufficiently small constants. Then with probability at least $1 - d^{-\Omega(C)}$, a smoothed instance $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$ satisfies the following for all $i, j \in [d - L + 1]$:

(a). $\text{ed}(x_{[i:i+L-1]}, x_{[j:j+L-1]}) \geq c_a \cdot \min\{pL, |j - i|\}$, and similarly for y .

(b). If $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) \geq \beta L$ for some $\frac{c_c}{32} < \beta < 1$, then $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq c_b \frac{\beta}{\log 2/\beta} \cdot pL$.

(c). Let $k^* = \text{match}_{A^*}(i, L)$, then

$$\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq \min\{c_c \cdot pL, c_c \cdot |j - k^*| - 2 \text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[k^*:k^*+L-1]}^*)\}.$$

Furthermore, if $|j - k^*| \geq L$, then $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq c_c \cdot pL$.

Proof. It suffices to prove these bounds for fixed $i, j \in [d - L + 1]$, because the lemma follows by a union bound, when $C > 0$ is sufficiently large.

We start by proving part (a). Consider the case when $|j - i| \geq L$. Since the corresponding blocks $x_{[i:i+L-1]}$ and $x_{[j:j+L-1]}$ do not overlap, they are perturbed independently of each other. We will use the following observation: for every collection of events $\mathcal{E}, A_1, \dots, A_t$,

$$\Pr[\cup_i A_i] \leq \Pr[\bar{\mathcal{E}}] + \sum_i \min\{\Pr[A_i], \Pr[A_i | \mathcal{E}]\}. \quad (2.1)$$

In particular, let \mathcal{E} be the event that at least $pL/4$ bits in $x_{[i:i+L-1]}$ are flipped by the perturbation. Note that $\Pr[\bar{\mathcal{E}}] \leq e^{-\Omega(pL)}$ by Chernoff bound. We shall use the notation $\hat{p} = \min\{p, \frac{1}{2}\}$ (the reason to consider \hat{p} is that the cases $p \leq \frac{1}{2}$ and $p > \frac{1}{2}$ require slightly different bounds).

Consider a potential alignment A between the two blocks $x_{[i:i+L-1]}$ and $x_{[j:j+L-1]}$, i.e. a map $A : [L] \rightarrow [L] \cup \{\perp\}$ that is monotonically increasing on $A^{-1}([L])$, and assume that $\text{cost}(A) = 2|A^{-1}(\perp)|$ equals αpL for a constant $0 < \alpha < 1/6$ to be determined later. Consider the mismatches under A between the unperturbed blocks namely,

$$M_A = \{k \in A^{-1}([L]) : x_{[i-1+k]}^* \neq x_{[j-1+A(k)]}^*\}. \quad (2.2)$$

Suppose first that $|M_A| \geq pL/16$. Then for A to be an actual alignment between the two blocks in x , for every mismatch $k \in M$, the perturbation must flip exactly one of the two relevant bits, which happens with probability $2 \cdot p/2 \cdot (1 - p/2) \leq \hat{p}$. Since these events are independent, in this case

$$\Pr \left[A \text{ is an alignment} \right] \leq \hat{p}^{pL/16}.$$

Now we consider the case when $|M_A| < pL/16$, and assume that the event \mathcal{E} occurs. Then the number of mismatches after perturbing only $x_{[i:i+L-1]}$, i.e. number of $k \in A^{-1}([L])$ such that $x_{[i-1+k]}^* \neq x_{[j-1+A(k)]}^*$, is at least $pL/4 - |M_A| - |A^{-1}(\perp)| \geq pL/16$. For A to be an actual alignment between the two perturbed blocks, all the corresponding positions $j - 1 + A(k)$ must also be flipped by the perturbation. Since each of these happens with probability $p/2$ and they are independent,

$$\Pr \left[A \text{ is an alignment} \mid \bar{\mathcal{E}} \right] \leq (p/2)^{pL/16} \leq \hat{p}^{pL/16}.$$

The number of potential alignments of cost αpL is exactly (as one needs to determine the unaligned positions in each block)

$$\binom{L}{\frac{1}{2}\alpha pL}^2 \leq \left(\frac{2e}{\alpha p}\right)^{\alpha pL} \leq \left(\frac{1}{\alpha^2 p}\right)^{\alpha pL}. \quad (2.3)$$

Finally we apply (2.1), and, by choosing $\alpha > 0$ to be a sufficiently small constant independent of p , we obtain that

$$\Pr[\underline{\text{ed}}(x_{[i:i+L-1]}, x_{[j:j+L-1]}) \leq \alpha pL] \leq e^{-\Omega(pL)} + \left(\left(\frac{1}{\alpha^2 p}\right)^{16\alpha} \cdot \hat{p}\right)^{pL/16} \cdot \alpha pL \leq e^{-\Omega(pL)}.$$

The above proof immediately extends to the case $|j - i| \geq L/4$ (the constant $1/4$ is arbitrary here). Indeed, consider in each block the initial segment of length $t = |j - i|$, which do not overlap. By the above argument, with high probability $\underline{\text{ed}}(x_{[i:i+t-1]}, x_{[j:j+t-1]}) \geq \Omega(pt) = \Omega(pL)$, implying a similar lower bound for the two blocks of length L .

Next we consider the remaining case, when $t = |j - i| < L/4$. Note that in this slightly harder case, the blocks $x_{[i:i+L-1]}$ and $x_{[j:j+L-1]}$ have a large overlap and thus we do not have the easy independence from before.

Assume without loss of generality that $i < j$. As before, consider a potential alignment $A : [L] \rightarrow [L] \cup \{\perp\}$ of cost $\alpha \cdot \min\{pL, t\}$ for a constant $0 < \alpha < 1/16$ to be determined later. Observe that for every $k \in A^{-1}([L])$, we have $|k - A(k)| \leq \frac{1}{2} \text{cost}(A) \leq \frac{1}{2}\alpha t$, thus $i - 1 + k \neq j - 1 + A(k)$, and in particular these two positions are perturbed independently of each other.

Define M_A as in Eqn. 2.2, and consider the case when $|M_A| \geq pL/64$. Then for A to be an actual alignment, for every $k \in M_A$ we must have $x[i - 1 + k] = x[j - 1 + A(k)]$, i.e., exactly one of the two relevant bits must be flipped by the perturbation. These events are not independent, but we can find at least $1/3$ of them that are independent (because every bit $x[l]$ appears in at most two such requirements). Thus in this case,

$$\Pr[A \text{ is an alignment}] \leq \hat{p}^{pL/192}.$$

Next suppose that $|M_A| < pL/64$. Partition the interval $[i : i + L - 1]$ into subintervals of length $t/2$, and take every fourth subinterval starting from the first one, namely $I = [i : i + \frac{t}{2} - 1] \cup [i + 2t : i + 2.5t - 1] \cup \dots$. We define \mathcal{E} to be the event that at least $pL/16$ positions in I are flipped by the perturbation. Notice that the definition of this event does not depend on A , and that by a Chernoff bound, $\Pr[\bar{\mathcal{E}}] \leq e^{-\Omega(pL)}$. As before, we shall assume that the event \mathcal{E} occurs. Observe that if $k \in A^{-1}([L])$ and $i - 1 + k \in I$ then $j - 1 + A(k) \notin I$ (because the difference between these two positions is $j - i + A(k) - k$ which falls in the range $[t - \frac{1}{2}\alpha t, t + \frac{1}{2}\alpha t]$). After conditioning on the outcomes of the perturbations inside I (only), the number of such k for which $x_{[i-1+k]} \neq x_{[j-1+A(k)]}^*$ is at least $pL/16 - |M_A| - |A^{-1}(\perp)| \geq pL/64$. For A to be an actual alignment between the two perturbed blocks, all the corresponding positions $j - 1 + A(k)$ must also be flipped by the perturbation. Since each of these happens with probability $p/2$ and they are independent,

$$\Pr[A \text{ is an alignment} \mid \bar{\mathcal{E}}] \leq (p/2)^{pL/64} \leq \hat{p}^{pL/64}.$$

The number of potential alignments of cost $\alpha \cdot \min\{pL, t\} \leq \alpha pL$ is at most $\left(\frac{1}{\alpha^2 p}\right)^{\alpha pL}$ by Eqn. 2.3. Hence, applying (2.1) and choosing $\alpha > 0$ to be a sufficiently small constant independent of p , we have $\Pr[\underline{\text{ed}}(x_{[i:i+L-1]}, x_{[j:j+L-1]}) \leq \alpha \cdot \min\{pL, t\}] \leq e^{-\Omega(pL)}$. This completes the proof of part (a).

Before continuing to parts (b) and (c), we prove the following claim, used for both parts. It is a variant of the argument from above for alignments between blocks of x and y .

Claim 2.3. Fix $i, j \in [d - L + 1]$, and let $\beta > 0$ and $\alpha \leq O(\frac{\beta}{\log 1/\beta})$. Suppose that, for any potential alignment $A : [L] \rightarrow [L] \cup \{\perp\}$ between $x_{[i:i+L-1]}$ and $y_{[j:j+L-1]}$ of cost at most αpL , there exists a set $S \subset [L]$, of size $|S| = \beta L$, such that for every $k \in S \setminus A^{-1}(\perp)$, we have that $j + A(k) - 1 \neq A^*(i + k - 1)$ (that is, A and A^* map to different positions in y for positions in S).

Then, $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \leq \alpha pL$ with probability at most $e^{-\Omega(\beta pL)}$.

Proof. Fix a potential alignment A of cost αpL . First, we can pick a subset $\bar{S} \subset S$, such that all events $x[i - 1 + k] \stackrel{?}{=} y[j - 1 + A(k)]$ for $k \in \bar{S} \setminus A^{-1}(\perp)$ are independent. Formally, \bar{S} is such that $A^*(i - 1 + \bar{S} \setminus A^{-1}(\perp)) \cap (j + A(\bar{S}) \setminus A^{-1}(\perp)) = \emptyset$. The maximal such set has size $|\bar{S}| \geq |S|/2$.

Define $M_A \subseteq \bar{S}$ to be those positions in \bar{S} which are non-matching positions under A in x^*, y^* :

$$M_A = \{k \in A^{-1}([L]) \cap \bar{S} : x_{[i-1+k]}^* \neq y_{[j-1+A(k)]}^*\}.$$

First, suppose $M_A \geq \frac{p}{32}\beta L$. Then, for each $k \in M_A$, the event $x_{[i-1+k]} = y_{[j-1+A(k)]}$ happens only with probability at most \hat{p} . Since all these events are independent (due to that fact that $M_A \subseteq \bar{S}$), we conclude that A is a valid alignment with probability at most $\hat{p}^{|M_A|} \leq \hat{p}^{p\beta L/32}$.

Now suppose $M_A < \frac{p}{32}\beta L$. Define \mathcal{E} to be the event that there are at least $\frac{p}{8}\beta L$ positions $k \in \bar{S}$ that are flipped: $x[i + k - 1] \neq x^*[i + k - 1]$. Note that $\Pr[\mathcal{E}] \leq e^{-\Omega(p\beta L)}$ by Chernoff bound. Now we condition on the event \mathcal{E} . Consider the positions $k \in \bar{S} \setminus A^{-1}(\perp)$ such that $x[i+k-1] \neq y^*[j+A(k)-1]$; the number of such positions is at least $\frac{p}{8}\beta L - |M_A| - |A^{-1}(\perp)| \geq \frac{p}{16}\beta L$. For each such position k , the event $x[i+k-1] = y[j+A(k)-1]$ happens with probability $p/2 \leq \hat{p}$. Furthermore, all such events are independent, even after we condition on \mathcal{E} , and thus A is a valid alignment with probability at most $\hat{p}^{p\beta L/16}$.

The number of alignments of cost αpL is at most $(\frac{1}{\alpha^2 p})^{\alpha pL}$ by Eqn. (2.3). Finally, applying (2.1), we obtain that

$$\Pr[\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \leq \alpha pL] \leq e^{-\Omega(p\beta L)} + \hat{p}^{p\beta L/16} \cdot \alpha pL \cdot (\frac{1}{\alpha^2 p})^{\alpha pL}.$$

The conclusion follows by choosing $\alpha = O(\frac{\beta}{\log 1/\beta})$. This completes the proof of Claim 2.3. \square

Part (b) follows easily from the above claim. In particular, suppose $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) = \beta L$ for some $\beta > 0$. Let

$$U = \{k \in [L] : A^*(i + k - 1) \notin [j : j + L - 1]\}$$

be the set of positions in $x_{[i:i+L-1]}^*$ not matched into $y_{[j:j+L-1]}^*$ under A^* . Note that $|U| = \beta L$. Then just apply the Claim 2.3 with $S = U$.

We now proceed to proving part (c). Suppose $\beta L = \text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[k^*:k^*+L-1]}^*) \geq L/4$. Then $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) \geq L/4$ and the claim results by part (b).

Now, consider $\beta L \leq L/4$. If $|j - k^*| \geq L/2$, then $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) \geq L/2 - L/4$, and the inequality follows by part (b).

Otherwise, when $|j - k^*| < L/2$, we use the triangle inequality to deduce that

$$\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq \text{ed}(y_{[k^*:k^*+L-1]}, y_{[j:j+L-1]}) - \text{ed}(x_{[i:i+L-1]}, y_{[k^*:k^*+L-1]}).$$

Thus, by part (a), and using the fact that $\text{ed}(x_{[i:i+L-1]}, y_{[k^*:k^*+L-1]}) \leq 2 \text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[k^*:k^*+L-1]}^*) = 2\beta L$, we have

$$\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq c_a \cdot \min\{pL, |j - k^*|\} - 2\beta L.$$

If $\beta L < \frac{c_a}{4}pL$, then we obtain that

$$\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq \min\{\frac{c_a}{2}pL, c_a \cdot |j - k^*| - 2\beta L\}.$$

For the other case, if $\frac{c_a}{4}pL \leq \beta L \leq L/4$, the statement is non-vacuous only when $|j - k^*| \geq \frac{2}{c_c} \cdot \beta L$ (note that we are still to choose c_c). Assuming this last condition, we prove that $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) \geq \Omega(pL)$.

We now want to show that we can apply Claim 2.3. Wlog, suppose $j > k^*$. Consider a potential alignment $A : [L] \rightarrow [L] \cup \{\perp\}$ between $x_{[i:i+L-1]}$ and $y_{[j:j+L-1]}$ of cost αpL for $0 < \alpha < \frac{\beta}{4p}$. Let S be the set of matchings in A between $x[i : i + L - 1]$ and $y[k^* : k^* + L - 1]$:

$$S = \{z \in [i : i + L - 1] : A(z) \in [1 : k^* - j + L]\}$$

We note that, for each $z \in S$, if we consider $x[i + z - 1]$, then A and A^* cannot map it to the same symbol in y . Formally, for each $z \in S$, we have that $j + A(z) - 1 \neq A^*(i + z - 1)$ since $j + A(z) - 1 \geq j + (z - i) - \alpha pL/2$ while $A^*(z) \leq k^* + (z - i) + \beta L$ or $A^*(z) = \perp$. Moreover, among positions $k \in [k^* + L : j + L - 1]$, only for at most βL such k 's we have $(A^*)^{-1}(k) \in [i : i + L - 1]$. Thus, any potential alignment A can have at most βL alignments that are the same as in A^* . We are now in position to apply Claim 2.3, thus completing the proof. \square

3 Near-Linear Time Distance Estimation

Our first algorithm is guaranteed to give a correct answer for any input strings, but has an improved runtime for smoothed inputs, coming from a distribution $\text{Smooth}_p(x^*, y^*, A^*)$.

Theorem 3.1. *For every $\varepsilon > 0$ and $p > 0$ there is a deterministic algorithm that, given as input two strings $x, y \in \{0, 1\}^d$, approximates $\text{ed}(x, y)$ within factor $O(\frac{1}{\varepsilon p} \log \frac{1}{\varepsilon p})$, and on a p -smoothed instance, with high probability its running time is $O(d^{1+\varepsilon})$.*

We will need three lemmas, the first two of which do not deal directly with smoothed instances and may be useful in other scenarios as well.

Lemma 3.2. *Consider a bipartite graph $G = ([d], [d], E)$, and call two edges $(i, j) \in E$ and $(k, l) \in E$ intersecting if $(i - k)(j - l) \leq 0$. Then a maximum-cardinality subset of non-intersecting edges can be found in time $O(d + |E| \log d)$ by reducing the problem to Patience Sorting.*

Proof of Lemma 3.2. Construct a string z of length $|E|$ via the following procedure. Start with an empty string. For each node i , $i = 1 \dots d$, we append to the end of z the list of characters $(j_1, -i), \dots, (j_k, -i)$ where $j_1 > j_2 > \dots > j_k$ are the neighbors of i : $\{j_1, \dots, j_k\} = \{j \in [d] \mid (i, j) \in E\}$. Notice that j_1, \dots, j_k are appended in the *decreasing* order.

It should now be clear that the longest increasing subsequence of x gives a maximum subset of non-intersecting edges (the order of symbols $(j, -i)$ is the lexicographic one). More precisely an increasing sequence $(j_1, -i_1) < \dots < (j_l, -i_l)$ forms a non-intersecting set $(i_1, j_1), (i_2, j_2), \dots, (i_l, j_l)$ and vice-versa.

The string z has length E , and thus, using Patience Sorting (or just straightforward dynamic programming), we can find the longest increasing subsequence of z in $O(|z| \log |z|) = O(|E| \log d)$ time. \square

The next lemma gives some necessary conditions on an optimal alignment between two strings. In this section we will only need the first two parts, which are easier to state, but in section 4 we will use all the four parts.

Lemma 3.3. *Fix an optimal alignment A between two strings $x, y \in \{0, 1\}^d$. Let $L \in [d]$ divide d . Partition x into successive blocks of length L , denoted $(X_i)_{i=1}^{d/L}$, and let $Y_i = \text{match}_A(X_i)$. Let V be the set of i for which $\text{ed}_A(X_i, Y_i) < L$. Then the following holds.*

(a). $\sum_{i \in [d/L]} \text{ed}_A(X_i, Y_i) \leq 2 \text{ed}(x, y)$.

(b). For $\varepsilon > 0$, let $B_\varepsilon = \{i \in [d/L] : \text{ed}(X_i, Y_i) > \varepsilon L\}$. Then $|B_\varepsilon| \leq \frac{4}{\varepsilon L} \cdot \text{ed}(x, y)$.

(c). For $i \in [d/L]$, let s_i be the starting position of Y_i . Then for all $i, i' \in V$ and $i < i'$, we have

$$s_{i'} - s_i \geq L - \text{ed}_A(X_i, Y_i) - \text{ed}_A(X_{i'}, Y_{i'}).$$

(d). For $i \in [d/L]$, let S_i be the positions in $y_{[s_i : s_i + L - 1]}$ that appear also in some block $y_{[s_{i'} : s_{i'} + L - 1]}$ for $i' \neq i$. Then $\sum_{i \in [d/L]} |S_i| \leq 2 \text{ed}(x, y)$.

Proof. For $i \in [d/L]$, let MIN_i and MAX_i , respectively, be the positions of the first and last aligned symbol in X_i , i.e., $MIN_i = \min\{j \in [iL - i + 1 : iL] \mid A(j) \in [d]\}$ and similarly for MAX_i . It could be that MIN_i and MAX_i are undefined, when $A(j) = \perp$ for all $j \in [iL - L + 1 : iL]$, in which case, abusing the notion, we define $A(MIN_i) = 0$ and $A(MAX_i) = -1$. Let u_i^x be the number of unaligned positions in $X_i = x_{[iL - L + 1 : iL]}$, i.e. $u_i^x = |\{j \in [iL - L + 1 : iL] \mid A(j) = \perp\}|$. Also, let u_i^y be the number of unaligned positions in $y_{[A(MIN_i) : A(MAX_i)]}$. If MIN_i, MAX_i are undefined, then set $u_i^x = L$ and $u_i^y = 0$.

If $A(MAX_i) - A(MIN_i) < L$, then $\text{ed}_A(X_i, Y_i) = u_i^x$. If $A(MAX_i) - A(MIN_i) \geq L$, then $\text{ed}_A(X_i, Y_i) \leq u_i^x + u_i^y$. Observing that each of $\sum_i u_i^x$ and $\sum_i u_i^y$ is bounded by $\text{ed}(x, y)$ proves part (a).

To prove part (b), notice that if $\text{ed}(X_i, Y_i) > \varepsilon L$, then $\text{ed}_A(X_i, Y_i) \geq \frac{1}{2} \text{ed}(X_i, Y_i) > \frac{1}{2} \varepsilon L$. By previous part, there could be at most $\frac{4}{\varepsilon}$ such blocks, and thus the claimed bound on the size of B_ε .

For part (c), note that, since $A(MAX_i) < A(MIN_{i'})$, we have

$$\begin{aligned} s_i &\leq A(MAX_i) + 1 - (L - \text{ed}_A(X_i, Y_i)) \\ &\leq A(MIN_{i'}) - L + \text{ed}_A(X_i, Y_i). \end{aligned}$$

We also observe that $s_{i'} \geq A(MIN_{i'}) - \text{ed}_A(X_{i'}, Y_{i'})$. Taking $s_{i'} - s_i$ implies the inequality.

Finally, we prove part (d). For $i \in [d/L]$, define r'_i to be the number of $j \in [s_i : s_i + L - 1]$ such that $j \notin [A(MIN_i) : A(MAX_i)]$. Then we argue that $2 \sum_{i \in V} r'_i \geq \sum_{i \in V} |S_i|$. We charge all contribution in $\sum_{i \in V} |S_i|$ to the positions $j \in [s_i : s_i + L - 1]$ with $j \notin [A(MIN_i) : A(MAX_i)]$. Any such position j contributes at most twice to the sum $\sum_{i \in V} |S_i|$: at most once in $|S_i|$ and once in $|S_k|$ for $k \in V$ such that $j \in [A(MIN_k) : A(MAX_k)]$.

We conclude that $\sum_{i \in [d/L]} |S_i| \leq 2 \text{ed}(x, y)$. We upper bound r'_i by u_i^x . Indeed, if $A(MAX_i) - A(MIN_i) \geq L$, then $r'_i = 0$. If $A(MAX_i) - A(MIN_i) \leq L - 1$, then $r'_i = L - 1 - (A(MAX_i) - A(MIN_i)) \leq u_i^x$. Also, for $i \notin V$, we have that $|S_i| \leq L = u_i^x$. Thus, $\sum_{i \in V} |S_i| \leq 2 \sum_{i \in V} r'_i + L \cdot (d/L - |V|) \leq 2 \sum u_i^x \leq 2 \text{ed}(x, y)$. \square

Lemma 3.4. *Let $C > 1$ and $0 < c' < 1$ be sufficiently large and sufficiently small constants, respectively, and let $L = \frac{C}{p} \log d$. Let A^* be a maximum-length alignment between $x^*, y^* \in \{0, 1\}^d$. Then for every $i \in [d]$ there is $j_i^* \in [d]$ such that, for $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$, with probability at least $1 - d^{-2}$, for all j with $|j - j_i^*| > L$ we have $\text{ed}(x_{[i:i+L-1]}, y_{[j:j+L-1]}) > c'pL$.*

Proof. Take $j_i^* = \text{match}_{A^*}(x_{[i:i+L-1]}^*)$. If $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j_i^*:j_i^*+L-1]}^*) < L/4$, then for all j with $|j - j_i^*| > L$ we have $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) \geq L - L/4$. Otherwise, for all j we have $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[j:j+L-1]}^*) \geq L/4$. In both cases the conclusion results by applying Lemma 2.2(b). \square

We now proceed to complete the proof of Theorem 3.1.

Proof of Theorem 3.1. We use as a building block a near neighbor (NN) data structure under edit distance, defined as follows. Preprocess a database of m strings each of length L , so that given a query string, the algorithm returns *all* database strings at distance $\leq \varepsilon L$ from the query. We will construct such data structure at the end, and for now assume it can be implemented with preprocessing $P(m, L)$ and query time $Q(m, L) + O(|\text{output}|)$, where *output* is the list of points reported by the query.

Let $C > 1$ and L be as in Lemma 3.4 and assume $\varepsilon < c'p$. Our algorithm proceeds in two stages. The first one uses the NN data structure to find, for each position in x , a few ‘‘candidate matches’’ in y , presumably including the correct match (under optimal alignment) for a large fraction of positions in x . The second stage views the candidate matches between positions in x and in y as the edge-set E of a bipartite graph and applies the algorithm from Lemma 3.2, thereby reconstructing an alignment.

Let us describe the algorithm in more detail. The first stage builds an NN data structure on all the substrings of length L in y . Then, it partitions x into successive blocks $x_{[iL-L+1:iL]}$, and for each such block, queries the NN data structure to identify all blocks in y that are within distance εL . For each such block in y , collect all the character matches between the two blocks, i.e., every zero in the block in x with every zero in the block in y , and same for ones. Let E be the resulting list of all candidate matches. The second stage simply applies Lemma 3.2 to this list E to retrieve an alignment between x and y . The reported approximation to $\text{ed}(x, y)$ is then twice the cost of this alignment.

Next we argue the correctness of the algorithm. Consider an optimal alignment A between x and y . Lemma 3.3 guarantees that for all but $4 \text{ed}(x, y)/\varepsilon L$ blocks from x , there exists a corresponding block $y_{[s_i:s_i+L-1]}$ at distance $\leq \varepsilon L$. Since the algorithm detects all pairs of blocks at distance $\leq \varepsilon L$, the lemma implies that all but $O(\frac{1}{\varepsilon}) \text{ed}(x, y)$ of aligned pairs from the alignment A will appear in the list of candidate matches. The algorithm will then compute an alignment A' that has at least $d - O(\frac{1}{\varepsilon}) \text{ed}(x, y)$ aligned pairs. Concluding, the algorithm will output a distance D such that $\text{ed}(x, y) \leq D \leq O(\frac{1}{\varepsilon}) \text{ed}(x, y)$.

Next we show that, with high probability, the running time of the algorithm is $O(dL \log d + P(d, L) + \frac{d}{L} \cdot Q(d, L))$. Indeed, by Lemma 3.4, for each query block $x_{[iL-L+1:iL]}$, only blocks $y_{[j:j+L-1]}$ for $|j - j_{iL-L+1}^*| \leq L$ can be at distance εL . Thus, for each position in $x_{[iL-L+1:iL]}$, we have at most $3L$ candidate matches, hence $|E| \leq O(dL)$. We can now conclude that the first stage runs in $O(P(d, L) + d/L \cdot (Q(d, L) + L^2))$, and the second stage runs in $O(|E| \log d) = O(dL \log d)$.

Finally, it remains to describe the NN data structure. We achieve $P(m, L) = m \log m \cdot 2^{L \cdot O(\varepsilon \log 1/\varepsilon)}$ preprocessing and $Q(m, L) = O(L)$ query time. The data structure simply prepares all answers in advance: for each string σ in the database and every string τ at edit distance $\leq \varepsilon L$ from

σ , store the pair (σ, τ) in a trie keyed by τ (we can also use a hash table if we allow a randomized algorithm). To query a string q , the algorithm accesses the trie using q as the key, and for every pair (η, q) returned by the trie, it reports the string η . Recall that a trie with t strings of length L , has query time $O(L)$, and preprocessing time $O(tL \log t)$. Thus, $Q(m, L) \leq O(L)$ and since there are at most $\binom{2L}{\varepsilon L}^3$ strings at edit distance $\leq \varepsilon L$ from a given string,

$$P(m, L) \leq O(m \log m \cdot \left(\frac{2L}{\varepsilon L}\right)^4 \cdot L) \leq m \log m \cdot 2^{L \cdot O(\varepsilon \log(1/\varepsilon))}.$$

The overall running time becomes $d^{1+O(p^{-1}\varepsilon \log(1/\varepsilon))}$ for $O(1/\varepsilon)$ approximation. To complete the proof, apply the above to $\varepsilon' = \Theta(\varepsilon p / \log \frac{1}{p\varepsilon})$. The resulting running time is $d^{1+\varepsilon}$ and the approximation is $O(1/\varepsilon') = O(\frac{1}{\varepsilon p} \log \frac{1}{\varepsilon p})$. \square

4 Sublinear Time Distance Estimation

We now present a sublinear time algorithm that estimates the edit distance of a smoothed instance (x, y) within a constant factor. The precise guarantees are stated in the following theorem. As before, we assume that $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$, where A^* is the optimal alignment between two strings $x^*, y^* \in \{0, 1\}^d$, and $0 < p \leq 1$.

Theorem 4.1. *For every $\varepsilon > 0$ there is a randomized algorithm that, given as input $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$, approximates $\text{ed}(x, y)$ within factor $O(\frac{1}{\varepsilon p} \log \frac{1}{\varepsilon p})$ in time $O(d^{1+\varepsilon} / \sqrt{\text{ed}(x, y)})$, with success probability at least $1 - d^{-2}$ (over the randomness in the smoothing operation and the algorithm's coins).*

The high-level approach is to map the smoothed instance (x, y) to a pair of permutations (P, Q) , such that the edit distance between x and y is approximately equal to the Ulam distance between P and Q . We can then estimate the Ulam distance between P and Q using an off-the-shelf sublinear algorithm for estimating Ulam distance. Specifically, we use the following algorithm of [AIK09].

Theorem 4.2 ([AIK09]). *There exists a randomized algorithm that, given access to two permutations P, Q of length d , approximates $\text{ed}(P, Q)$ within a constant factor in time $\tilde{O}(d / \sqrt{\text{ed}(P, Q)})$, with success probability at least $2/3$.*

We remark that this algorithm is based on adaptive sampling, i.e. query positions depend on the outcome of earlier queries. As mentioned earlier, a direct application of this theorem implies a much weaker version of Theorem 4.1, with approximation factor $O(\log d)$, by employing the mapping of [CK06, Theorem 3.1], which views each block (with overlaps) in x or y as a symbol in a large alphabet $\{0, 1\}^L$. Thus, the main challenge we face is to obtain $O(1)$ -approximation.

A key observation is that the algorithm in Theorem 4.2 (for Ulam distance estimation) works exactly the same way regardless of any relabeling of the symbols used in P, Q . More precisely, when the algorithm queries one character, say $P[i]$, the algorithm uses only the information of whether $P[i]$ is identical to a previously queried character $Q[j]$, and vice versa. Other than the value of j , and the information whether such j exists, the name of the read symbol is not important.⁴ This observation can be leveraged in the following way: if the algorithm is about to query $P[i]$, and

⁴In fact, it is plausible that every algorithm for Ulam distance estimation can be transformed into one satisfying the stated property, without any loss in approximation factor and query complexity.

the matching character $Q[j]$ (i.e. position j such that $P[i] = Q[j]$) was not queried yet, then we may relabel this unread symbol, changing both $P[i]$ and $Q[j]$ to an arbitrary other symbol that does not appear anywhere at all. (Of course, such $Q[j]$ might not exist or might not be queried at all by a sublinear algorithm.) For the sake of analysis (but not in the algorithm) we may further assume (again by relabeling symbols) that Q is a fixed permutation, say the identity ($Q[j] = j$ for all $j \in [d]$). In the sequel, the permutations P, Q will always be of length d and over the alphabet $\Sigma = [2d]$.

We shall construct P, Q (from x, y) based on the following principle. Let A be an alignment between x and y , say of near-optimal cost $O(\text{ed}(x, y))$. Then we can construct P (while Q is the identity) so that A is the *optimal* alignment between P and Q , as follows: set $P[i] = Q[A(i)]$ whenever $A(i) \in [d]$, and set $P[i] = d + i$ whenever $A(i) = \perp$. For our purpose, A has to be computable “on the fly”. More precisely we require that, for every two queried positions i, j in P, Q respectively, we can determine whether $A(i) = j$ by querying only $x[i]$ and $y[j]$, possibly together with a small neighborhood around these positions; in particular, it is independent of the rest of the strings x, y . We term this property *locality*, and ensuring it is the main technical part of our proof. We note that, for worst-case strings (x, y) , constructing a near-optimal alignment A that satisfies the locality property seems hard; for a smoothed instance, on the other hand, we show this is possible, largely due to Lemma 2.2 and Lemma 3.3. In our presentation below, we will not construct the alignment A explicitly, but instead show how to construct P directly. The actual construction of P, Q will differ from the above description in that it will actually work with whole blocks rather than single characters.

4.1 Block Structure Lemma

Lemma 4.3.

We state a technical lemma that provides some structural properties of the edit distance between smoothed strings. These properties have a local nature, and will be useful later when we design our reduction, to prove, roughly speaking, that our local operations result in a correct (global) edit distance. As before, we assume that $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$, where A^* is the optimal alignment between two strings $x^*, y^* \in \{0, 1\}^d$, and $0 < p \leq 1$.

Lemma 4.4. *Consider a smoothed instance $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$. Let $L = \frac{C}{p} \log d$ for a sufficiently large constant $C > 0$. Partition x into successive blocks of length L , denoted $X_1, X_2, \dots, X_{d/L}$, and let $Y_k = \text{match}(X_k)$ for $k \in [d/L]$. For sufficiently small $\varepsilon > 0$, let $M = \{k \in [d/L] : \text{ed}(X_k, Y_k) \leq \varepsilon p L\}$. Then, with probability $\geq 1 - d^{-\Omega(C)}$, we have:*

- (a). $d/L - |M| \leq \frac{4}{\varepsilon p L} \cdot \text{ed}(x, y)$.
- (b). $\sum_{k \in M} \text{ed}(X_k, Y_k) \leq 4 \cdot \text{ed}(x^*, y^*)$.
- (c). Let S_k , for $k \in [d/L]$, be the set of positions in Y_k that appear also in some block $Y_{k'}$ for $k' \in M \setminus \{k\}$; then $\sum_{k \in M} |S_k| \leq O(1) \cdot \text{ed}(x^*, y^*)$.
- (d). The starting positions of blocks Y_k , for $k \in M$, are in a strictly increasing order, and moreover the distance between two consecutive such positions is $> L/2$.

Proof. For part (a), the upper bound on $d/L - |M|$ follows from Lemma 3.3(b) applied to the strings x, y , since $B_{\varepsilon p}$ in the language of that lemma is precisely $[d/L] \setminus M$.

We now prove part (b). Let X_k^*, Y_k^* denote the blocks from x^*, y^* that correspond to (i.e. have the same positions as) blocks X_k, Y_k , respectively. Let $s_k^* = \text{match}_{A^*}(X_k^*)$ for all $k \in [d/L]$. The following inequality is immediate.

$$\text{ed}(X_k, Y_k) \leq \text{ed}(X_k, y_{[s_k^*:s_k^*+L-1]}^*) \leq \text{ed}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*). \quad (4.1)$$

Combining Eqn. (4.1) with Lemma 3.3(a) and using an immediate relation between ed and ed_{A^*} , we obtain

$$\sum_{k \in [d/L]} \text{ed}(X_k, Y_k) \leq 2 \sum_{k \in [d/L]} \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*) \leq 4 \cdot \text{ed}(x^*, y^*).$$

We now prove part (c). Before bounding $\sum_{k \in M} |S_k|$ itself, we prove the following two claims, where j_k denotes the starting position of Y_k . Let $0 < c_c < 1$ be the constant from Lemma 2.2.

Claim 4.5. *With probability at least $1 - d^{-\Omega(C)}$, for all $k \in M$:*

$$|j_k - s_k^*| \leq \frac{3}{c_c} \cdot \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*). \quad (4.2)$$

Proof. Fix $k \in M$ and notice that, by Lemma 2.2 (c), with high probability,

$$\text{ed}(X_k, Y_k) \geq \min\{c_c \cdot pL, c_c \cdot |j_k - s_k^*| - \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*)\}.$$

For $k \in M$ and assuming $\varepsilon < \min\{c_b, c_c\}$, the minimum must be attained by the second term, hence $|j_k - s_k^*| \leq \frac{1}{c_c} (\text{ed}(X_k, Y_k) + \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*))$. Finally, by Eqn. (4.1),

$$\text{ed}(X_k, Y_k) \leq \text{ed}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*) \leq 2 \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*),$$

and altogether this proves Claim 4.5. \square

Now let S_k^* be the positions in $y_{[s_k^*:s_k^*+L-1]}^*$ that also appear in some block $y_{[s_{k'}^*:s_{k'}^*+L-1]}^*$ for some other $k' \in M \setminus \{k\}$. Lemma 3.3(d) tells us that $\sum_{k \in M} |S_k^*| \leq 2 \text{ed}(x^*, y^*)$.

Claim 4.6. *With probability at least $1 - d^{-\Omega(C)}$, we have $\sum_{k \in M} |S_k| \leq \sum_{k \in M} (S_k^* + 2|s_k^* - j_k|)$.*

Proof. Consider $k, k' \in M$ with $k < k'$. Observe that $\text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*)$ and $\text{ed}_{A^*}(X_{k'}^*, y_{[s_{k'}^*:s_{k'}^*+L-1]}^*)$ must be at most $\frac{c_c}{32} \cdot L$, or otherwise, by Lemma 2.2(b), with high probability, $\text{ed}(X_k, Y_k)$ or $\text{ed}(X_{k'}, Y_{k'})$ respectively is at least $c_b \cdot pL$, and thus k or k' respectively is not in M .

We can now apply Lemma 3.3(c) to obtain that

$$s_{k'}^* - s_k^* \geq L - \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*) - \text{ed}_{A^*}(X_{k'}^*, y_{[s_{k'}^*:s_{k'}^*+L-1]}^*)$$

and this, together with Claim 4.5, yields

$$\begin{aligned} j_{k'} - j_k &\geq s_{k'}^* - s_k^* - |j_k - s_k^*| - |j_{k'} - s_{k'}^*| \\ &\geq L - \frac{4}{c_c} \cdot \left[\text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*) + \text{ed}_{A^*}(X_{k'}^*, y_{[s_{k'}^*:s_{k'}^*+L-1]}^*) \right]. \end{aligned}$$

Concluding, since $\text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*)$ and $\text{ed}_{A^*}(X_{k'}^*, y_{[s_{k'}^*:s_{k'}^*+L-1]}^*)$ are at most $\frac{c_c}{32} \cdot L$, we have

$$j_{k'} - j_k > L/2. \quad (4.3)$$

Thus, every position in Y_k for $k \in M$ can appear in at most one other block $Y_{k'}$ for $k' \in M \setminus \{k\}$. It not difficult to see that $\sum_{k \in M} |S_k| \leq \sum_{k \in M} (S_k^* + 2|s_k^* - j_k|)$, since every position in S_k either contributes also to S_k^* , or to $|s_k^* - j_k|$, or to some $|s_{k'}^* - j_{k'}|$, for $k' \in M$, and furthermore the contributions of the same type are all distinct. \square

Combining all the above, namely applying Claim 4.6, then Lemma 3.3(d) and Claim 4.5, and finally Lemma 3.3(a), we have

$$\sum_{k \in M} |S_k| \leq \sum_{k \in M} (S_k^* + 2|s_k^* - j_k|) \leq 2 \text{ed}(x^*, y^*) + O(1) \sum_{k \in M} \text{ed}_{A^*}(X_k^*, y_{[s_k^*:s_k^*+L-1]}^*) \leq O(\text{ed}(x^*, y^*)).$$

Part (d) of the lemma follows directly from Eqn. (4.3). \square

4.2 Reducing Smoothed Instances to Ulam Instance

Next we show how to efficiently translate a smoothed instance of edit distance into an instance of Ulam's distance, while distorting the distance by only a constant factor. As mentioned earlier, for the sake of analysis we may set Q to be the identity permutation, and construct P as a function of x and y . The lemma below defines P in its entirety, while ensuring the locality property: that every character in P can be computed from local information. (As we shall see later, the algorithm uses this locality property to compute "on the fly" the same P, Q , up to relabeling of the symbols.)

The basic idea appears simple. First, we partition P into blocks of length $L = O(\frac{1}{p} \log d)$. Then, each such block $x_{[kL-L+1:kL]}$ is matched to its closest block in y , say $y_{[l:l+L-1]}$, and then $P_{[kL-L+1:kL]}$ is defined in a simple way that depends only on $Q_{[l:l+L-1]}$ and on $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]})$, and satisfies $\text{ed}(P_{[kL-L+1:kL]}, Q_{[l:l+L-1]}) = \text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]})$. This reduction preserves the edit distance locally (at the block level), although it is not clear it is true also globally (for the entire strings). We indeed prove the latter, i.e. that $\text{ed}(P, Q)$ approximates $\text{ed}(x, y)$, using the technical machinery developed in Lemma 4.4.

The main challenge we face in implementing this basic idea is that characters may repeat in P , because the blocks we match against in y may overlap with each other. A straightforward fix to this issue could be to change these repetitions to completely new symbols (distinct symbols that do not appear in Q). This fix increases $\text{ed}(P, Q)$, although, as we show, only by a small factor. Unfortunately, this fix also introduces dependencies between different blocks in P , violating the locality requirement. We thus refine this fix by going through two smaller transformations of P , which reduces the dependencies of a position to only the nearby blocks.

Lemma 4.7 (Reduction to Ulam). *Fix $\varepsilon > 0$, let $(x, y) \in \text{Smooth}_p(x^*, y^*, A^*)$ and $L = \frac{C}{p} \log d$ for a sufficiently large constant $C > 0$. Then there exist two permutations P and $Q = (1, 2, \dots, d)$ such that, with probability $\geq 1 - d^{-\Omega(C)}$, the following holds.*

- **Distance:** $\Omega(1) \cdot \text{ed}(x, y) \leq \text{ed}(P, Q) \leq O(\frac{\log 1/p}{p} + \frac{1}{\varepsilon p}) \cdot \text{ed}(x, y)$; and
- **Locality:** For all $k \in [d/L]$, $j \in [kL - L + 1 : kL]$, and $s_k = \text{match}(x_{[kL-L+1:kL]})$:
 1. $P[j]$ can be computed from only s_k , $x_{[kL-2L+1:kL]}$, and $y_{[s_k-6L:s_k+L-1]}$, in time $O(L^3)$. $P[j]$ is either $P[j] \in [d]$ or $P[j] = d + j$.
 2. For all $l \in [d - L + 1]$ s.t. $|l - s_k| \geq 2L$, we have $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]}) > \Omega(pL)$. Furthermore, if $P[j] \in [d]$ then necessarily $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]}) \leq \varepsilon pL$.

Proof. In the sequel, we say that position $j \in [d]$ (in P) is *invalidated* if it is set to the symbol $d+j$. All other positions will be set to symbols in the range $[d]$. Recall that the alphabet is $\Sigma = [2d]$ and that Q is the identity, hence invalidated positions in P match no character in Q . We first give a complete description of the construction of P , and then prove its properties (distance and locality).

The permutation P is constructed by first defining a string P^1 , then invalidating some positions to obtain P^2 , and then invalidate more positions to obtain the final P . The intermediate strings P^1 and P^2 might not be permutations. We now describe these three stages and a preceding setup stage.

Setup: Let $L = \frac{C}{p} \log d$ be the block length. Partition x into d/L blocks of length L , denoted X_k , and for each $k \in [d/L]$ let $Y_k = \text{match}(X_k)$. Let s_k be the starting position of Y_k and let $c_k = \text{ed}(X_k, Y_k)$. Let $M = \{k \in [d/L] : \text{ed}(X_k, Y_k) \leq \varepsilon p L\}$.

P¹: For every $k \in M$, set $P^1_{[kL-L+1:kL]}$ to be equal to the block $Q_{[s_k:s_k+L-1]}$, except that the first c_k symbols are invalidated (thus ensuring $\text{ed}(P^1_{[kL-L+1:kL]}, Q_{[s_k:s_k+L-1]}) = c_k$). For each $k \in [d/L] \setminus M$, invalidate the entire block $P^1_{[kL-L+1:kL]}$.

P²: Let $F \subseteq M$ be the following set. For each $k \in M$, $k > 1$, put k into F if **(i)** $k-1 \notin M$; or **(ii)** $k-1 \in M$ and $s_k - s_{k-1} > 2L$. We obtain P^2 by invalidating all blocks $P^1_{[kL-L+1:kL]}$ with $k \in F$.

P: Invalidate all positions $j \in [d]$ for which the symbol $P^2[j]$ occurs previously in P^2 to the left of the position j . That is, for each symbol, we invalidate all its occurrences except the very first one.

We proceed to prove the distance property, using two claims that provide a lower bound and an upper bound on $\text{ed}(P, Q)$, respectively.

Claim 4.8. *With high probability, $\text{ed}(x, y) \leq 6 \cdot \text{ed}(P, Q)$.*

Proof. First observe that $\text{ed}(P^1, Q) \leq \text{ed}(P, Q)$ because invalidating some positions can only increase the edit distance. We proceed to show that $\text{ed}(x, y) \leq 6 \cdot \text{ed}(P^1, Q)$. Fix an optimal alignment \tilde{A} between P^1 and Q , and construct an alignment A between x and y as follows. For each $k \in [d/L]$, consider the block X_k . If $k \notin M$ (i.e. $c_k > \varepsilon p L$), then the corresponding block in P^1 has only invalidated positions, which cannot be aligned to Q , hence the same alignment is valid for x on these blocks. Otherwise, by construction, $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]}) = c_k = \text{ed}(P^1_{[kL-L+1:kL]}, Q_{[s_k:s_k+L-1]})$. Let t_k be the number of non-aligned positions in the k^{th} block of P^1 under \tilde{A} . Clearly, $t_k \geq c_k$ since k^{th} block of P^1 has c_k invalidated positions that cannot match any position in Q . We construct A by aligning the corresponding k^{th} block in x against only the middle $L - 2t_k$ symbols in $y_{[s_k:s_k+L-1]}$, in the best possible way. The number of unaligned positions in $x_{[kL-L+1:kL]}$ is at most $c_k + 2t_k \leq 3t_k$, and thus $\text{ed}(x, y) \leq 2 \text{ed}_A(x, y) \leq 2 \sum_k 3t_k \leq 6 \text{ed}_{\tilde{A}}(P^1, Q) \leq 6 \text{ed}(P^1, Q)$.

It remains to show that A is a valid alignment. Consider $k, k' \in M$ with $k < k'$. Observe that, while we are not guaranteed that $s_k + L - 1 < s_{k'}$, by definition of t_k we must have $(s_k + L - 1) - t_k < s_{k'} + t_{k'}$. Since our construction of A is limited to the middle $L - 2t_k$ symbols in each $y_{[s_k:s_k+L-1]}$, we get that A is indeed monotonically increasing on $A^{-1}(\perp)$. \square

Claim 4.9. *With high probability, $\text{ed}(P, Q) \leq O\left(\frac{\log(1/p)}{p} + \frac{1}{\varepsilon p}\right) \text{ed}(x, y)$.*

Proof. First we argue that the non-invalidated positions of P (in any of the three stages) form an increasing sequence, and thus $\frac{1}{2} \text{ed}(P, Q)$ is upper bounded by the number of the invalidated positions. Note that we are precisely in the conditions of Lemma 4.4. We use the notation from that lemma for the rest of this proof, and assume that the whp event holds. The lemma says that the starting positions of Y_k , for $k \in M$, are strictly increasing and moreover increase by $> L/2$ each time. Thus, after the invalidations, a block $P_{[kL-L+1:kL]}$ is either completely invalidated (if $k \notin M \setminus F$), or its invalidated positions form a contiguous sequence at the beginning of the block. Furthermore, in the latter case, the symbol in the position of kL is smaller than any non-invalidated symbol in $P_{[kL+1:d]}$. Thus all the non-invalidated position of P form an increasing sequence.

We now upper bound the number of invalidated positions in the construction of P^1 , in the transformation to P^2 , and in the transformation to P . The number of positions invalidated in the construction of P^1 is $L \cdot (d/L - |M|) + \sum_{k \in M} c_k$. The number of positions invalidated in the transformation to P^2 is at most $L \cdot |F|$. The number of positions invalidated in the transformation to P is at most $\sum_k |S_k|$, because the number of positions in $P_{[kL-L+1:kL]}$ invalidated in this step is at most $|S_k|$. Lemma 4.4 bounds all these quantities, except for $|F|$.

We now bound $L \cdot |F|$. Notice that each $k \in F$ corresponds either to a block $k-1 \in [d/L] \setminus M$ (case (i) in the definition of F) or to some block of length L strictly between the blocks Y_{k-1} and Y_k (case (ii)). Positions appearing in blocks of the latter type do not belong to any Y_k for $k \in M$, and thus their number is at most $L \cdot (d/L - |M|)$ plus the number of positions that appear in two blocks $Y_k, Y_{k'}$ for distinct $k, k' \in M$. Thus, $L \cdot |F| \leq 2L \cdot (d/L - |M|) + \sum_k |S_k|$.

Concluding, using Lemma 4.4, we get that, with high probability,

$$\begin{aligned} \frac{1}{2} \text{ed}(P, Q) &\leq \sum_{k \in M} c_k + L \cdot (d/L - |M|) + L \cdot |F| + \sum_{k \in M} |S_k| \\ &\leq \sum_{k \in M} \text{ed}(X_k, Y_k) + 3L \cdot (d/L - |M|) + 2 \sum_{k \in M} |S_k| \\ &\leq O(1) \cdot \text{ed}(x^*, y^*) + O\left(\frac{1}{\varepsilon p}\right) \text{ed}(x, y). \end{aligned}$$

Futhermore, using Theorem 2.1, we conclude that $\text{ed}(P, Q) \leq O\left(\frac{\log(1/p)}{p} + \frac{1}{\varepsilon p}\right) \text{ed}(x, y)$. \square

We proceed to proving the locality property. In our language, j is inside the block X_k and we need to prove that $P[j]$ depends only on the blocks X_k, Y_k together with regions of $6L$ positions preceding them. The algorithm for computing $P[j]$ follows the description of the construction of P .

Specifically, we compute $P[j]$ in a local manner as follows. For $Y_k = \text{match}(X_k)$ starting at s_k in y , we set $P[j] = s_k + j - (kL - L + 1)$, unless it is invalidated according to the following procedure, in which case $P[j] = d + j$. If $\text{ed}(X_k, Y_k) > \varepsilon pL$, then j is invalidated (stage 1 invalidation). Also, if $kL - L + 1 + \text{ed}(X_k, Y_k) > j$, then j is invalidated (stage 1 invalidation). In case $k > 1$, we consider the block X_{k-1} and define s'_{k-1} to be the position $s \in [s_k - 4L : s_k]$ that minimizes $\text{ed}(X_{k-1}, y_{[s:s+L-1]})$. If $\text{ed}(X_{k-1}, y_{[s'_{k-1}:s'_{k-1}+L-1]}) > \varepsilon pL$ or if $s_k - s'_{k-1} > 2L$, then j is invalidated (stage 2 invalidation). Next, we take

$$s'_{k-2} = \text{argmin}_{s \in [s'_{k-1}-4L:s'_{k-1}]} \text{ed}(X_{k-2}, y_{[s:s+L-1]}).$$

If $k = 2$ or $\text{ed}(X_{k-2}, y_{[s'_{k-2}:s'_{k-2}+L-1]}) \leq \varepsilon pL$ and $s'_{k-1} - s'_{k-2} \leq 2L$, and also if $j - (kL - L + 1) < (s'_{k-1} + L) - s_k$, then j is invalidated (stage 3 invalidation).

The correctness of this algorithm follows from the claim that $P[j]$ is invalidated in the above procedure if and only if it is invalidated in the original three-stages construction. We argue this claim next. It is easy to see that stage 1 invalidations are correct. Next, suppose $k > 1$ and j is not invalidated in stages 1 or 2 of the original construction, which happens if $k - 1 \in M$ and $s_k - s_{k-1} \leq 2L$. In this case, our algorithm will obtain $s'_{k-1} = s_{k-1}$, and thus $P[j]$ will not be invalidated in stage 2 of the above algorithm. Conversely, if $k - 1 \notin M$, then we have $\text{ed}(X_{k-1}, y_{[s'_{k-1}:s'_{k-1}+L-1]}) > \varepsilon pL$, or else $s_k - s_{k-1} > 2L$, and then $s'_{k-1} \leq \max\{s_{k-1}, s_k - 4L + (2L - 1)\} < s_k - 2L$ (namely, if $s_{k-1} < s_k - 4L$, then $\text{ed}(X_{k-1}, y_{[s:s+L-1]}) > \varepsilon pL$ for all $s \geq s_k - 2L$ by Lemma 2.2(c)). In both cases, the algorithm invalidates $P[j]$. Finally, if $P[j]$ was not invalidated in stages 1 or 2, $P[j]$ can be invalidated in the third stage if Y_k intersects with Y_{k-1} (when $k - 1 \in M \setminus F$). To check the intersection with Y_{k-1} , the algorithm checks additionally that $k - 2 \in M$ (unless $k = 2$) and $s_{k-2} - s_{k-1} \leq 2L$ (implying $k - 1 \in M \setminus F$). If this condition passes, then we invalidate $P[j]$ iff symbol $P[j] = s_k + j - (kL - L + 1)$ is in the intersection of Y_k and Y_{k-1} (note that the matching symbol $P[j - L - s_k + s_{k-1}]$ in $P_{[kL-2L+1:kL-L]}$ cannot have been invalidated in stages 1 and 2).

We now prove the second part. By construction, $P[j] = d + j$ if $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]}) > \varepsilon pL$. Now, let $k^* = \text{match}_{A^*}(i, L)$, and let c_c be the constant from Lemma 2.2. If $\text{ed}_{A^*}(x_{[i:i+L-1]}^*, y_{[k^*:k^*+L-1]}^*) \geq \frac{c_c}{4}L$, then, for all $l \in [d - L + 1]$, we have $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]}) = \Omega(pL)$ by Lemma 2.2(b), and we reach the desired conclusion. Next, suppose that $\text{ed}_{A^*}(x_{[kL-L+1:kL]}^*, y_{[k^*:k^*+L-1]}^*) < \frac{c_c}{4}L$. Then, for every $l \in [d - L + 1]$ such that $|l - k^*| \geq L$, we have that $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]}) = \Omega(pL)$ by Lemma 2.2(c). If $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]}) \leq \varepsilon pL$, then $|s_k - k^*| \leq L$, and thus, for all $l \in [d - L + 1]$ s.t. $|l - s_k| \geq 2L$, we have $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]}) > \Omega(pL)$. But if $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]}) > \varepsilon pL$ then there is nothing to prove since, by definition, s_k is the l that minimizes $\text{ed}(x_{[kL-L+1:kL]}, y_{[l:l+L-1]})$. \square

4.3 Sublinear time algorithm for smoothed instances

We now describe the sublinear algorithm for the smoothed instance $\text{Smooth}_p(x^*, y^*, A^*)$ and thus prove Theorem 4.1. The algorithm basically performs a reduction to a similar problem (distance estimation) under the Ulam metric, and solving the latter using the algorithm of [AIK09] in a black-box fashion.

Proof of Theorem 4.1. We will use the sublinear algorithm from [AIK09] as a black box. We call their algorithm \mathcal{A} and note that \mathcal{A} makes $\tilde{O}\left(\frac{d}{\sqrt{\text{ed}(P,Q)}}\right)$ queries to P, Q and has the same running time, while succeeding with constant probability. For completeness, we note that the algorithm \mathcal{A} reduces the problem to several *decision version* problems of distance estimation, and then solves the decision version problem. More precisely, the decision version is, for a given threshold $R \in [d]$, to decide whether the distance is $\text{ed}(P, Q) \leq R$ or $\text{ed}(P, Q) > \alpha R$ for approximation factor $\alpha = O(1)$. The algorithm of [AIK09] for the decision version runs in time $\tilde{O}\left(\frac{d}{\sqrt{R}}\right)$.

We would like to run \mathcal{A} on the permutations P, Q obtained from applying Lemma 4.7 to our input (x, y) . Since we cannot afford to compute the entire P , our reduction will generate on the fly (and feed them into \mathcal{A}) two permutations that are equivalent to P and Q , up to a relabeling of the symbols. As explained at the beginning of section 4, the algorithm \mathcal{A} is independent of the actual names of the symbols, hence its output is invariant under this relabeling.

We proceed to describing our reduction, which can be viewed as a data structure that has random access to x and y , and provides a random access interface to the permutations P and Q

(modulo relabeling). This data structure will be used by the algorithm \mathcal{A} . Let L be defined as in Lemma 4.7, and assume that the high-probability event described in the lemma holds.

Our reduction keeps for each of P and Q two data structures, one to keep track of the relabeling and one to keep track of the blocks. Let \hat{P} store the relabeling of P , namely, $\hat{P}(i)$ for a position $i \in [d]$ represents the new symbol given to $P[i]$ by the relabeling, or the value \perp if the position i has not been queried in P yet. We assume \hat{P} is implemented so as to support fast inverse search, i.e. given a symbol $a \in [2d]$ it can report $\hat{P}^{-1}(a)$. Let T_P be a trie (or another data structure implementing a dictionary) that stores the substrings $x_{[kL-L+1:kL]}$ for all $k \in [d/L]$ such that at least one of the positions among $[kL - L + 1 : kL]$ was queried in P . We define \hat{Q} for Q analogously to \hat{P} ; note that if $P[i]$ and $Q[j]$ have already been queried, then $\hat{P}(i) = \hat{Q}(j)$ if and only if $P[i] = Q[j] = j$. Finally, the trie T_Q for Q stores (i) the substrings $y_{[l:l+L-1]}$ for all l such that there is at least one queried position $j \in [d]$ in Q with $|l - j| \leq L$, as well as (ii) all length L strings within edit distance εpL from such $y_{[l:l+L-1]}$.

The query to a position $P[i]$ works as follows. If $\hat{P}(i) \neq \perp$, return $\hat{P}(i)$. Otherwise, add the substring $X_k = x_{[kL-L+1:kL]}$, where $i \in [kL-L+1 : kL]$, into the trie T_P (unless it is already present there). Then check whether X_k is present in the trie T_Q ; if it is not, then assign a new symbol to $P[i]$, update $\hat{P}(i)$ accordingly, and return this symbol $\hat{P}(i)$. Suppose now that X_k is present in T_Q , i.e. it matches a string at edit distance at most εpL from a block $y_{[l:l+L-1]}$. Then apply the locality algorithm from Lemma 4.7, and compute $P[i]$. More precisely, compute s_k , with $|l - s_k| \leq 2L$, that minimizes $\text{ed}(x_{[kL-L+1:kL]}, y_{[s_k:s_k+L-1]})$. Note that this s_k indeed satisfies $s_k = \text{match}(x_{[kL-L+1:kL]})$ by Lemma 4.7, Locality-2 property. Applying the algorithm from Lemma 4.7, we compute $P[i]$ and store the value in \hat{P} . If $P[i] \notin [d]$, assign a new symbol to $P[i]$ and add it to \hat{P} . Otherwise (i.e. $P[i] \in [d]$ and thus $P[i] = Q[j] = j$ for some j), check whether the position $P[i]$ has been queried in Q by checking whether $\hat{Q}(P[i]) \neq \perp$. If indeed $\hat{Q}(P[i]) \neq \perp$, then update accordingly $\hat{P}(i) = \hat{Q}(P[i])$. And if $\hat{Q}(P[i]) = \perp$, then assign a new symbol to $P[i]$ and update $\hat{P}(i)$ accordingly. In the end, return the symbol $\hat{P}(i)$.

The query to a position $Q[j]$ is almost analogous, with the obvious modifications to account for the asymmetry in the two tries T_P and T_Q . Namely, if \hat{Q} already holds a symbol for $Q[j]$, return it. Otherwise, add every substring $y_{[l:l+L-1]}$, where $|l - j| \leq L$, together with all the length L strings at edit distance $\leq \varepsilon pL$, into the trie T_Q . For each added string, check whether the string is present in T_P ; if it is, i.e. the added string matches some X_k stored in T_P , then compute all of $P_{[kL-L+1:kL]}$ using Lemma 4.7, and if for any of the computed position i we have $P[i] = j$ and $\hat{P}(i) \neq \perp$, then the new symbol for $Q[j]$ is that symbol, i.e. set $\hat{Q}(j) = \hat{P}(i)$. If, at the end, $\hat{Q}(j)$ is still \perp , then we assign $Q[j]$ with a new symbol, and update $\hat{Q}(j)$ accordingly. Either way, return the symbol $\hat{Q}(j)$.

The correctness of the algorithm then follows immediately from the Locality part of Lemma 4.7. In particular, essentially by construction, \hat{P}, \hat{Q} form a consistent relabeling of P, Q , respectively, although a partial one in the sense that some of the values are replaced by \perp .

To obtain the stated bounds on query complexity and runtime, we note that the overhead on each query to P or Q is $O(L)$ queries to x, y and $O(L^{O(1)} d^{O(\varepsilon \log 1/\varepsilon)})$ time, where $d^{O(\varepsilon \log 1/\varepsilon)}$ is an upper bound on the number of strings at distance $\leq \varepsilon pL$ from any single string (of length L). To obtain the claimed dependence on ε , we replace the ε used in the above algorithm with $\varepsilon' = O(\varepsilon / \log \frac{1}{\varepsilon})$. \square

5 Conclusions

It seems challenging to obtain a distance estimation algorithm whose smoothed running time is quasi-linear, i.e. $d \cdot \log^{O(1)} d$, or whose approximation is independent of the smoothing parameter p at the expense of only $O(1/p)$ increase in the runtime. Perhaps more important is to extend the smoothed analysis framework to other problems, such as nearest neighbor search (or pattern matching). One may hope to match the $O(\log \log d)$ approximation that was recently obtained for the Ulam metric [AIK09].

Acknowledgments.

We thank Dick Karp for useful discussions at an early stage of this research.

References

- [AGM⁺90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *J. of Molecular Biology*, 215(3):403–410, 1990.
- [AIK09] A. Andoni, P. Indyk, and R. Krauthgamer. Overcoming the ℓ_1 non-embeddability barrier: Algorithms for product metrics. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 865–874, 2009.
- [AK07] A. Andoni and R. Krauthgamer. The computational hardness of estimating edit distance. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 724–734. IEEE, October 2007.
- [AO09] A. Andoni and K. Onak. Approximating edit distance in near-linear time. 2009. To appear in STOC’09.
- [BEK⁺03] T. Batu, F. Ergün, J. Kilian, A. Magen, S. Raskhodnikova, R. Rubinfeld, and R. Sami. A sublinear algorithm for weakly approximating edit distance. In *Proceedings of the Symposium on Theory of Computing*, pages 316–324, 2003.
- [BES06] T. Batu, F. Ergün, and C. Sahinalp. Oblivious string embeddings and edit distance approximations. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 792–801, 2006.
- [BJKK04] Z. Bar-Yossef, T. S. Jayram, R. Krauthgamer, and R. Kumar. Approximating edit distance efficiently. In *Proceedings of the Symposium on Foundations of Computer Science*, pages 550–559, 2004.
- [BS95] A. Blum and J. Spencer. Coloring random and semi-random k -colorable graphs. *J. Algorithms*, 19(2):204–234, September 1995.
- [CK06] M. Charikar and R. Krauthgamer. Embedding the ulam metric into ℓ_1 . *Theory of Computing*, 2(11):207–224, 2006.
- [FK01] U. Feige and J. Kilian. Heuristics for semirandom graph problems. *J. Comput. Syst. Sci.*, 63(4):639–673, 2001.

- [FM97] A. Frieze and C. McDiarmid. Algorithmic theory of random graphs. *Random Structures and Algorithms*, 10(1-2):5–42, 1997.
- [GP06] S. Gollapudi and R. Panigrahy. A dictionary for approximate string search and longest prefix search. In *15th ACM international conference on Information and knowledge management*, pages 768–775. ACM, 2006.
- [MP80] W. J. Masek and M. Paterson. A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.*, 20(1):18–31, 1980.
- [MTL02] B. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.
- [Nav01] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.
- [NBYST01] G. Navarro, R. Baeza-Yates, E. Sutinen, and J. Tarhio. Indexing methods for approximate string matching. *IEEE Data Engineering Bulletin*, 24(4):19–27, 2001. Special issue on Text and Databases. Invited paper.
- [ST03] D. A. Spielman and S.-H. Teng. Smoothed analysis: Motivation and discrete models. In *WADS, Lecture Notes in Computer Science*. Springer-Verlag, 2003.
- [ST04] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004.
- [WF74] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.