

# Approximating the Geometric Edit Distance\*

Kyle Fox<sup>†</sup>

Xinyi Li<sup>‡</sup>

September 9, 2020

## Abstract

Edit distance is a measurement of similarity between two sequences such as strings, point sequences, or polygonal curves. Many matching problems from a variety of areas, such as signal analysis, bioinformatics, etc., need to be solved in a geometric space. Therefore, the geometric edit distance (GED) has been studied. In this paper, we describe the first strictly sublinear approximate near-linear time algorithm for computing the GED of two point sequences in constant dimensional Euclidean space. Specifically, we present a randomized  $O(n \log^2 n)$  time  $O(\sqrt{n})$ -approximation algorithm. Then, we generalize our result to give a randomized  $\alpha$ -approximation algorithm for any  $\alpha \in [\sqrt{\log n}, \sqrt{n/\log n}]$ , running in time  $O(n^2/\alpha^2 \log n)$ . Both algorithms are Monte Carlo and return approximately optimal solutions with high probability.

## 1 Introduction

Ordered sequences are frequently studied objects in the context of similarity measurements, because sequence alignment plays a vital role in trajectory comparison and pattern recognition. As a consequence, several metrics have been developed to measure the similarity of two sequences, e.g., Fréchet distance, dynamic time warping, and their variations. Geometric edit distance, a natural extension of the string metric to geometric space, is the focus of this paper. This concept is formally introduced by Agarwal *et al.* [2]; however, a similar idea (extending string edit distance to a geometric space) has been applied in other ways during the past decade. Examples include an  $l^p$ -type edit distance for biological sequence comparison [3], ERP (Edit distance with Real Penalty) [4], EDR (Edit Distance on Real sequence) [5], TWED (Time Warping Edit Distance) [6] and a matching framework from Swaminathan *et al.* [7] motivated by computing the similarity of time series and trajectories. See also a survey by Wang *et al.* [8].

**Problem statement.** Geometric Edit Distance (GED) is the minimum cost of any matching between two geometric point sequences that respects order along the sequences. The cost includes a constant penalty for each unmatched point.

Formally, let  $P = \langle p_1, \dots, p_m \rangle$  and  $Q = \langle q_1, \dots, q_n \rangle$  be two point sequences in  $\mathbb{R}^d$  for some constant  $d$ . A monotone *matching*  $\mathcal{M}$  is a set of index pairs  $\{(i_1, j_1), \dots, (i_k, j_k)\}$  such that the first elements  $i$  (respectively, second elements  $j$ ) are distinct and for any two elements  $(i, j)$  and  $(i', j')$  in  $\mathcal{M}$ ,  $i < i'$  if  $j < j'$ .

---

\*A preliminary version of this work appeared in the Proceedings of the 30th International Symposium on Algorithms and Computation [1]. Most of the work was done while the second author was a student at the University of Texas at Dallas.

<sup>†</sup>The University of Texas at Dallas, [kyle.fox@utdallas.edu](mailto:kyle.fox@utdallas.edu)

<sup>‡</sup>The University of Utah, [xin\\_yi.li@utah.edu](mailto:xin_yi.li@utah.edu)

We call every unmatched point a *gap point*. Let  $\Gamma(\mathcal{M})$  be the set of all gap points. The *cost* of  $\mathcal{M}$  is defined as

$$\delta(\mathcal{M}) = \sum_{(i,j) \in \mathcal{M}} \text{dist}(p_i, q_j) + \rho(\Gamma(\mathcal{M})) \quad (1)$$

where  $\text{dist}(p, q)$  is the distance between points  $p$  and  $q$  (i.e. the Euclidean norm), and  $\rho(\Gamma(\mathcal{M}))$  is a function of all gap points, which is known as a *gap penalty function*. The use of gap points and the gap penalty function allows us to recognize good matchings even in the presence of outlier points. The distance is sensitive to scaling, so, we may only match point pairs that are sufficiently close together. In this paper, we use a linear gap function. That is to say,  $\rho(\Gamma(\mathcal{M})) = |\Gamma(\mathcal{M})| \cdot \ell$ , where  $\ell$  is a constant parameter called the *gap penalty*. Without loss of generality, we may assume  $\ell = 1$  when designing algorithms.

**Definition 1.** We denote the GED between two sequences  $P, Q$  as:

$$\text{GED}(P, Q) = \min_{\mathcal{M}} \delta(\mathcal{M}) = \min_{\mathcal{M}} \left( \sum_{(i,j) \in \mathcal{M}} \text{dist}(p_i, q_j) + |\Gamma(\mathcal{M})| \right)$$

where the minimum is taken over all monotone matchings.

**Prior work.** To simplify the presentation of prior work, we assume  $n \geq m$ . It is trivial to compute  $\text{GED}(P, Q)$  in  $O(mn)$  time by simply changing the cost of substitution in the original string edit distance (Levenshtein distance) dynamic programming algorithm [9]. Assuming  $k$  is the GED, we can achieve an  $O(nk)$  time algorithm by restricting our attention to the middle  $k$  diagonals of the dynamic programming table (see also Ukkonen [10]). There is a slightly subquadratic  $O(n^2/\log n)$  time algorithm [11] for the string edit distance, but it appears unlikely we can apply it directly to the geometric case. Accordingly, Gold and Sharir [12] proposed a different algorithm which can compute GED as well as the closely related dynamic time warping (DTW) distance in  $O(n^2 \log \log \log n / \log \log n)$  time in polyhedral metric spaces. Recent papers have shown conditional lower bounds for several sequence distance measures even with some restrictions. Assuming the Strong Exponential Time Hypothesis (SETH) [13], there is no  $O(n^{2-\delta})$  time algorithm for any constant  $\delta > 0$  for Fréchet distance [14], DTW over a constant size alphabet [15] or restricted to one-dimensional curves [16], and string edit distance on the binary alphabet [17, 16].<sup>1</sup> The latter of the above results implies the same conditional lower bound for GED, even assuming the sequences consist entirely of 0, 1-points in  $\mathbb{R}$ .

Due to these limitations and difficulties, many researchers have turned to approximation algorithms for these distances. In particular, much work has been done to explore approximate algorithms for Fréchet distance, DTW, and string edit distance [18, 19, 20, 2, 21, 22]. Bringmann and Mulzer [18] describe an  $\alpha$ -approximation algorithm for the discrete Fréchet distance that runs in time  $O(n \log n + n^2/\alpha)$  for any  $\alpha \in [1, n]$ . Chan and Rahmati [19] improved this running time to  $O(n \log n + n^2/\alpha^2)$ . Very recently, Kuszmaul [20] provided  $O(\alpha)$ -approximation algorithms with  $O((n^2/\alpha) \text{polylog } n)$  running times for edit distance over arbitrary metric spaces and DTW over well separated tree metrics. Another  $O(n^2/\alpha)$  time algorithm with an  $O(\alpha)$  approximation factor

<sup>1</sup>The (discrete) Fréchet and DTW distances are defined similarly to GED; however, they use one-to-many correspondences instead of one-to-one matchings, and they disallow the use of gap points. As in GED, DTW aims to minimize the sum of distances between corresponding points, while discrete Fréchet distance aims to minimize the maximum distance over corresponding points.

for *string* edit distance is to run Ukkonen’s [10]  $O(nk)$  time algorithm letting  $k$  be  $n/\alpha$ , and unmatched all characters if this algorithm cannot return the optimal matching. Similarly, we can obtain a different  $O(\alpha)$ -approximation algorithm for GED running in  $O(n^2/\alpha)$  time by making use of the  $O(nk)$  time exact algorithm mentioned above. There are many other approximation algorithms specialized for the string version of edit distance. In particular, an  $O(\sqrt{n})$ -approximation algorithm with linear running time can be acquired easily from an  $O(n+k^2)$  time exact algorithm [23]. More recent results include algorithms with  $(\log n)^{O(1/\varepsilon)}$  [21] and constant approximation ratios [22] with different running time tradeoffs. The latest result in this line of work is an  $O(n^{1+\varepsilon})$  time constant factor approximation algorithm for any  $\varepsilon > 0$  [24].

For GED specifically, a simple linear time  $O(n)$ -approximation algorithm was observed by Agarwal *et al.* [2]. In the same paper, they also offered a subquadratic time (near-linear time in some scenarios) approximation scheme on several well-behaved families of sequences. Using the properties of these families, they reduced the search space to find the optimal admissible path in the dynamic programming graph [2].

**Our results.** Inspired by the above applications and prior work, we commit to finding a faster approach to approximating GED between general point sequences while also returning the approximate best matching. Here, we give the first near-linear time algorithm to compute GED with a strictly sublinear approximation factor. We then generalize our result to achieve a tradeoff between the running time and approximation factor. Both of these algorithms are Monte Carlo algorithms, returning an approximately best matching with high probability.<sup>2</sup> To simplify our exposition, we assume the points are located in the plane (i.e.,  $d = 2$ ), and we assume the input sequences are the same length (i.e.,  $m = n$ ). We can easily extend our results to the unbalanced case, and our analysis implies that outside the plane, the running times and approximation ratios increase only by a factor polynomial in  $d$ .

**Theorem 1.** *Given two point sequences  $P$  and  $Q$  in  $\mathbb{R}^2$ , each with  $n$  points, there exists an  $O(n \log^2 n)$ -time randomized algorithm that computes an  $O(\sqrt{n})$ -approximate monotone matching for geometric edit distance with high probability.*

The intuitive idea behind this algorithm is very simple. We check if the GED is less than each of several geometrically increasing values  $g$ , each of which is less than  $O(\sqrt{n})$ . For each  $g$ , we transform the geometric sequences into strings using a randomly shifted grid and run the  $O(n+k^2)$  time exact algorithm for strings [23]. If the GED is less than  $g$ , then we get an  $O(\sqrt{n})$  approximate matching. If we never find a matching of cost  $O(\sqrt{n})$ , we simply leave all points unmatched as this empty matching is an  $O(\sqrt{n})$ -approximation for GED with high probability. We give the details for this  $O(\sqrt{n})$ -approximation algorithm in Section 2.

**Theorem 2.** *Given two point sequences  $P$  and  $Q$  in  $\mathbb{R}^2$ , each with  $n$  points, there exists an  $O(\frac{n^2}{\alpha^2} \log n)$ -time randomized algorithm that computes an  $O(\alpha)$ -approximate monotone matching for geometric edit distance with high probability for any  $\alpha \in [\sqrt{\log n}, \sqrt{n/\log n}]$ .*

The second algorithm uses similar techniques to the former, except we can no longer use the string edit distance algorithm as a black box. In particular, we cannot achieve our desired time-approximation tradeoff by just directly altering some parameters in our first algorithm. We discuss why in Section 3.1. To overcome these difficulties, we develop a constant-factor approximation algorithm to compute the GED of point sequences obtained by snapping points of the original

<sup>2</sup>We say an event occurs with high probability if it occurs with probability at least  $1 - \frac{1}{n^c}$  for some constant  $c > 0$ .

input sequences to grid cell corners. Our algorithm for these snapped points is based on the exact algorithm for string edit distance [23] but necessarily more complicated to handle geometric distances. So, we first introduce the  $O(n + k^2)$  time algorithm for strings in Section 4.1, and then describe our constant approximation algorithm for points in Section 4.2. We note that a key component of the string algorithm and our extension is a fast method for finding maximal length common substrings from a given pair of starting positions in two strings  $A$  and  $B$ . A similar procedure was needed in the discrete Fréchet distance approximation of Chan and Rahmati [19]. In Section 3, we present the algorithm for Theorem 2 using our constant approximation algorithm for snapped point sequences as a black box.

## 2 $O(\sqrt{n})$ -Approximation for GED

Recall that the main part of our algorithm is a decision procedure to check if the GED is less than a guess value  $g$ . There are two steps in this process:

1. Transform the point sequences into strings. To be specific, we partition nearby points into common groups and distant points into different groups to simulate the identical characters and different characters in the string version of edit distance.
2. Run a modification of the exact string edit distance algorithm of Landau *et al.* [23]. To better serve us when discussing geometric edit distance, we aim to minimize the number of insertions and deletions to turn  $S$  into  $T$  *only*; we consider substitution to have infinite cost. Details on this modified algorithm appear in Section 4.1.<sup>3</sup>

We explain how to transform the point sequences into strings in Section 2.1, and we analyze the approximation factor and running time in Sections 2.2 and 2.3.

For convenience, we refer to the string edit distance algorithm as  $SED(S, T, k)$ , where  $S$  and  $T$  are two strings with equal length. This algorithm will return a matching in  $O(n + k^2)$  time if the edit distance is at most  $k$ . We give an outline of our algorithm as Algorithm 1. Here,  $c$  is a sufficiently large constant, and we use  $\lg$  to denote the logarithm of base 2.

### 2.1 Transformation by a random grid

As stated above, the transformation technique should partition nearby points into common groups and distant points into different groups. We use a randomly shifted grid to realize this ideal; see Har-Peled [25] for an introduction to randomly shifted grids.

Recall  $P$  and  $Q$  lie in  $\mathbb{R}^2$ . We cover the space with a grid. Let the side length of each grid cell be  $\Delta$ , and let  $b$  be a vector chosen uniformly at random from  $[0, \Delta]^2$ . Starting from an arbitrary position, the grid shifts  $b_i$  units in each dimension  $i$ . For a point  $p$ , let  $id_{\Delta, b}(p)$  denote the cell which contains  $p$  in this configuration. We consider two points  $p_1 = (x_1, y_1)$ , and  $p_2 = (x_2, y_2)$  in this space.

**Lemma 1.** *We have  $P(id_{\Delta, b}(p_1) \neq id_{\Delta, b}(p_2)) \leq \min\{\frac{|x_1 - x_2| + |y_1 - y_2|}{\Delta}, 1\}$ .*

We use this observation in our algorithm and set  $\Delta = \frac{g}{\sqrt{n}}$  as each cell's side length.

---

<sup>3</sup>This variant of the string edit distance is really closer to the shortest common supersequence length of the strings rather than the traditional Levenshtein distance, but we stick with “edit distance” for simplicity.

---

**Algorithm 1:**  $O(\sqrt{n})$ -approximation algorithm for GED

---

**Input:** Point sequences  $P$  and  $Q$ **Output:** An approximately optimal matching for GED

```
1 if  $\sum_{i=1}^n \text{dist}(p_i, q_i) \leq 1$  then
2   | return matching  $\{(1, 1), \dots, (n, n)\}$ 
3 else
4   | for  $i := 0$  to  $\lceil \lg \sqrt{n} \rceil$  do
5     |  $g := 2^i$ 
6     | for  $j := 1$  to  $\lceil c \lg n \rceil$  do
7       | Transform  $P, Q$  to strings  $S, T$  using a randomly shifted grid
8       |  $\text{out} := \text{SED}(S, T, 12\sqrt{n} + 2g)$ 
9       | if  $\text{out} \neq \text{false}$  then
10        | | return out
11        | end
12      | end
13    | end
14    | return the empty matching
15 end
```

---

## 2.2 Time complexity

We claim the running time for Algorithm 1 is  $O(n \log^2 n)$ . Computing  $\sum_{i=1}^n \text{dist}(p_i, q_i)$  takes  $O(n)$  time. In the inner loop, the transformation operation (line 7) takes  $O(n)$  time assuming use of a hash table. The running time for  $\text{SED}(S, T, 12\sqrt{n} + 2g)$  is  $O(n)$  for  $g = O(\sqrt{n})$ . Summing over the outer loop and inner loop, the overall running time for Algorithm 1 is

$$\sum_{i=1}^{\lceil \lg \sqrt{n} \rceil} \sum_{j=1}^{\lceil c \lg n \rceil} O(n) = O(n \log^2 n).$$

## 2.3 Approximation ratio

In this section, we show that Algorithm 1 returns an  $O(\sqrt{n})$ -approximate matching with high probability.

**Notation.** For any monotone matching  $\mathcal{M}$ , we define  $C_S(\mathcal{M})$  as the cost of the corresponding edit operations for  $\mathcal{M}$  in the string case and  $C_G(\mathcal{M})$  to be  $\delta(\mathcal{M})$  as defined in (1) for the geometric case (as stated, there is no substitution operation in our modified string case). Let  $\mathcal{M}_G^*$  be the optimal matching for geometric edit distance, and let  $\mathcal{M}_S^*$  denote the optimal matching under the string configuration during a given iteration of the inner for loop. Our goal is to establish the relationship between  $C_G(\mathcal{M}_G^*)$  and  $C_G(\mathcal{M}_S^*)$ .

**Lemma 2.** *Consider an iteration of the outer for loop, and suppose  $\text{GED}(P, Q) \leq g$ . With a probability at least  $1 - \frac{1}{n^c}$ , at least one of the  $\lceil c \lg n \rceil$  iterations of the inner for loop will return a matching  $\mathcal{M}_S^*$  where  $C_S(\mathcal{M}_S^*) \leq 12\sqrt{n} + 2g$ .*

**Proof:** Let  $\mathcal{M}$  be a monotone matching, and let  $UM_{\mathcal{M}}$  be the set of unmatched indices. There are four subsets of pairs in  $\mathcal{M}$ :

- $OC_{\mathcal{M}}$ : In each pair, both indices' points fall into One cell, and the distance between the two points is less and equal to  $\Delta = \frac{g}{\sqrt{n}}$  (Close).
- $OF_{\mathcal{M}}$ : In each pair, both indices' points fall into One cell, and the distance between the two points is larger than  $\Delta = \frac{g}{\sqrt{n}}$  (Far).
- $DC_{\mathcal{M}}$ : In each pair, the indices' points are in Different cells, and the distance between the two points is less and equal to  $\Delta = \frac{g}{\sqrt{n}}$  (Close).
- $DF_{\mathcal{M}}$ : In each pair, the indices' points are in Different cells and the distance between the two points is larger than  $\Delta = \frac{g}{\sqrt{n}}$  (Far).

These sets are disjoint, so

$$\begin{aligned}
C_G(\mathcal{M}_G^*) &= |UM_{\mathcal{M}_G^*}| + \sum_{(i,j) \in OC_{\mathcal{M}_G^*}} \text{dist}(p_i, q_j) + \sum_{(i,j) \in OF_{\mathcal{M}_G^*}} \text{dist}(p_i, q_j) \\
&+ \sum_{(i,j) \in DC_{\mathcal{M}_G^*}} \text{dist}(p_i, q_j) + \sum_{(i,j) \in DF_{\mathcal{M}_G^*}} \text{dist}(p_i, q_j). \tag{2}
\end{aligned}$$

Recall that there is no substitution operation in our version of the string edit distance. Therefore, to better understand optimal matchings for string edit distance, we unmatched all the pairs in  $DC_{\mathcal{M}_G^*}$  and  $DF_{\mathcal{M}_G^*}$ , forming a new matching  $\mathcal{M}_G^{*'}$ . Points in one cell are regarded as identical characters while those in different cells are different characters. Therefore,

$$\begin{aligned}
C_S(\mathcal{M}_G^{*'}) &= |UM_{\mathcal{M}_G^*}| + 0 \cdot (|OC_{\mathcal{M}_G^*}| + |OF_{\mathcal{M}_G^*}|) + 2 \cdot (|DC_{\mathcal{M}_G^*}| + |DF_{\mathcal{M}_G^*}|) \\
&= |UM_{\mathcal{M}_G^*}| + 2 \cdot (|DC_{\mathcal{M}_G^*}| + |DF_{\mathcal{M}_G^*}|).
\end{aligned}$$

Observe that there are at most  $\frac{g}{g/\sqrt{n}} = \sqrt{n}$  pairs in  $DF_{\mathcal{M}_G^*}$  if  $C_G(\mathcal{M}_G^*) \leq g$ . Therefore,

$$\begin{aligned}
C_S(\mathcal{M}_S^*) &\leq C_S(\mathcal{M}_G^{*'}) \\
&= |UM_{\mathcal{M}_G^*}| + 2|DC_{\mathcal{M}_G^*}| + 2|DF_{\mathcal{M}_G^*}| \leq g + 2\sqrt{n} + 2|DC_{\mathcal{M}_G^*}| \tag{3}
\end{aligned}$$

For any two points  $p_i, q_j$ , let  $P_D(i, j)$  be the probability that  $p_i$  and  $q_j$  are assigned into different cells. From Lemma 1, we can infer  $P_D(i, j) \leq \frac{2\text{dist}(p_i, q_j)}{g/\sqrt{n}}$ .

Then,

$$\begin{aligned}
\mathbb{E}(|DC_{\mathcal{M}_G^*}|) &\leq \sum_{(i,j) \in DC_{\mathcal{M}_G^*}} P_D(i, j) \leq \sum_{(i,j) \in DC_{\mathcal{M}_G^*}} \frac{2\text{dist}(p_i, q_j)}{g/\sqrt{n}} \\
&\leq 2\sqrt{n}. \tag{4}
\end{aligned}$$

Therefore,

$$\mathbb{E}(C_S(\mathcal{M}_S^*)) \leq 6\sqrt{n} + g.$$

By Markov's inequality,

$$P[C_S(\mathcal{M}_S^*) \geq 12\sqrt{n} + 2g] \leq \frac{1}{2}.$$

In other words,  $SED(S, T, 12\sqrt{n} + 2g)$  will fail with probability at most  $\frac{1}{2}$  if  $GED(P, Q) \leq g$ . So, if we test  $SED(S, D, 12\sqrt{n} + 2g)$   $\lceil c \lg n \rceil$  times, at least one iteration will return a value if  $GED(P, Q) \leq g$  with a probability greater than or equal to

$$1 - \prod_1^{\lceil c \lg n \rceil} P[C_S(\mathcal{M}_S^*) \geq 12\sqrt{n} + 2g] \geq 1 - \prod_1^{\lceil c \lg n \rceil} \frac{1}{2} = 1 - \frac{1}{n^c}.$$

We conclude the proof of Lemma 2.  $\square$

According to Lemma 2, with high probability, we obtain a matching  $\mathcal{M}_S^*$  such that  $C_S(\mathcal{M}_S^*) \leq 12\sqrt{n} + 2g$  if  $GED(P, Q) \leq g$ .

We now consider  $C_G(\mathcal{M}_S^*)$ . Again,  $UM_{\mathcal{M}}$  is the set of unmatched indices for a matching  $\mathcal{M}$ . Observe, for all  $(i, j) \in \mathcal{M}_S^*$ , points  $p_i$  and  $q_j$  lie in the same grid cell. Therefore,  $dist(p_i, q_j) \leq \frac{\sqrt{2}g}{\sqrt{n}}$  if  $(i, j) \in \mathcal{M}_S^*$ . We have:

$$\begin{aligned} C_G(\mathcal{M}_S^*) &= |UM_{\mathcal{M}_S^*}| + \sum_{(i,j) \in \mathcal{M}_S^*} dist(p_i, q_j) \\ &\leq 12\sqrt{n} + 2g + n \cdot \left(\frac{\sqrt{2}g}{\sqrt{n}}\right) = 12\sqrt{n} + 2g + \sqrt{2}g\sqrt{n} \end{aligned} \quad (5)$$

If  $GED(P, Q) \leq \sqrt{n}$ , then, with high probability, we obtain a matching  $\mathcal{M}_S^*$  by the end of the outer for loop iteration where  $g \geq GED(P, Q) \geq \frac{1}{2}g$ . The cost of this matching is at most  $12\sqrt{n} + 2g + \sqrt{2}g\sqrt{n} = O(\sqrt{n})GED(P, Q)$ . The same approximation bound holds if  $GED(P, Q) > \sqrt{n}$ , whether or not we find a matching during the outer for loop. We conclude the proof of Theorem 1.

### 3 $O(\alpha)$ -Approximation for GED

We now discuss our  $O(\alpha)$ -approximation algorithm for any  $\alpha \in [\sqrt{\log n}, \sqrt{n/\log n}]$ . A natural approach for extending our  $O(\sqrt{n})$ -approximation is using the same reduction to string edit distance but letting the cell's side length be a variable depending on the approximation factor  $\alpha$ . We argue, however, that this approach does not lead to a good approximation.

#### 3.1 Flaws in modifying the $O(\sqrt{n})$ -approximation to achieve a tradeoff

Suppose we try to modify the  $O(\sqrt{n})$ -approximation algorithm by simply changing the side length of cells. Let  $\Delta_\alpha$  be the cell's side length which depends on the approximation factor  $\alpha$ . We need to obtain a matching  $\mathcal{M}_S^*$  with high probability such that  $C_G(\mathcal{M}_S^*) \leq g \cdot O(\alpha)$  during any iteration of the outer for loop with  $GED(P, Q) \leq g$ .

There can be at most  $n$  matched pairs in  $\mathcal{M}_S^*$ . Following (5), we need  $n \cdot \Delta_\alpha \leq g \cdot O(\alpha)$ , implying

$$\Delta_\alpha \leq O\left(\frac{g\alpha}{n}\right).$$

On the other hand, we have  $C_S(\mathcal{M}_S^*) \leq C_G(\mathcal{M}_S^*) \leq g \cdot O(\alpha)$  in our analysis. We then derived  $2\sqrt{n}$  in (3) as  $2\frac{g}{\Delta_\alpha}$ . We now need  $2\frac{g}{\Delta_\alpha} \leq g \cdot O(\alpha)$ , implying

$$\Delta_\alpha \geq \Omega\left(\frac{1}{\alpha}\right).$$

This is fine for  $\alpha = \sqrt{n}$  or for large values of  $g$ . But for small  $\alpha$  and small  $g$ , we cannot have both inequalities be true. Therefore, we require a different approach to obtain our desired approximation factor-running time tradeoff.

### 3.2 $O(\alpha)$ -approximation algorithm based on grid-snapping

**Grid-snapping.** Instead of simply grouping points by their different cells as in the  $O(\sqrt{n})$ -approximation algorithm, we snap points to the lower left corners of their respective grid cells. Let  $P' = \langle p'_1, \dots, p'_n \rangle$ ,  $Q' = \langle q'_1, \dots, q'_n \rangle$  be the sequences after grid-snapping. Let  $\Delta$  be the cell side length of the grid. We immediately obtain the following observation:

**Observation 1.** For any  $p_i$  and  $q_j$  from  $P$  and  $Q$ , respectively, we have  $\text{dist}(p'_i, q'_j) \leq \text{dist}(p_i, q_j) + 2\sqrt{2}\Delta$ . Moreover, if  $p_i$  and  $q_j$  are in different cells,  $\text{dist}(p'_i, q'_j) \geq \Delta$ .

We can then obtain our  $O(\alpha)$ -approximation algorithm by altering the bound in the outer loop and the test procedure of Algorithm 1. See Algorithm 2. Here,  $AGED(P', Q', k)$  attempts to Approximate  $GED(P', Q')$  given that  $P'$  and  $Q'$  have their points on the corners of the grid cells. If  $GED(P', Q') \leq k$ , then it returns an  $O(1)$ -approximate matching for  $GED(P', Q')$ . Otherwise, it either returns an  $O(1)$ -approximate matching or it returns false.

---

#### Algorithm 2: $O(\alpha)$ -approximation algorithm

---

**Input:** Point sequences  $P$  and  $Q$   
**Output:** An approximately optimal matching for GED

```

1 if  $\sum_{i=1}^n \text{dist}(p_i, q_i) \leq 1$  then
2   | return matching  $\{(1, 1), \dots, (n, n)\}$ 
3 else
4   | for  $i := 0$  to  $\lceil \lg \frac{n}{\alpha} \rceil$  do
5     |  $g := 2^i$ 
6     | for  $j := 1$  to  $\lceil c \lg n \rceil$  do
7       | Obtain  $P', Q'$  by doing grid-snapping to  $P, Q$  based on a randomly shifted grid
8       |  $out := AGED(P', Q', (4\sqrt{2} + 6)g)$ 
9       | if  $out \neq false$  then
10      | | return out
11      | end
12     | end
13   | end
14   | Return the empty matching
15 end

```

---

We describe how to implement  $AGED(P', Q', k)$  in Section 4.2. The running time of our implementation is  $O(n + \frac{k^2}{\Delta})$  where  $\Delta$  is the cell side length of the grid. We do grid snapping in  $O(n)$  time. For each  $g = 2^i$ , we use cells of side length  $\frac{g\alpha}{n}$  and set  $k$  to  $(4\sqrt{2} + 6)g$ , so the overall running time of our  $O(\alpha)$ -approximation algorithm is

$$O(n) + \sum_{i=0}^{\lceil \lg \frac{n}{\alpha} \rceil} \sum_{j=1}^{\lceil c \lg n \rceil} O(n + \frac{2^i n}{\alpha}) = \sum_{i=0}^{\lceil \lg \frac{n}{\alpha} \rceil} O(n \log n + \frac{2^i n}{\alpha} \log n) = O(n \log^2 n + \frac{n^2}{\alpha^2} \log n).$$



Notice that if  $\alpha < \sqrt{\log n}$ , the running time of our algorithm is  $O(n^2)$ . Thus, we could just run the classic  $O(n^2)$  dynamic programming algorithm if we need an approximation factor to be less than  $\sqrt{\log n}$ . On the other hand, the  $O(n \log^2 n)$  in the running time is asymptotically insignificant if  $\alpha \leq \sqrt{n/\log n}$ . As a result, the running time is  $O(\frac{n^2}{\alpha^2} \log n)$  for any  $\alpha \in [\sqrt{\log n}, \sqrt{n/\log n}]$ .

### 3.3 Proof of correctness

The analysis for the  $O(\alpha)$ -approximation algorithm is similar to the first algorithm. First, we introduce some additional notation to that used in Section 2.3.

Let  $C_{GS}(\mathcal{M})$  be the cost of any monotone matching  $\mathcal{M}$  using distances between the grid-snapped points of  $P'$  and  $Q'$ . Let  $\mathcal{M}_{GS}^*$  be the optimal matching for  $P'$  and  $Q'$ , i.e.,  $C_{GS}(\mathcal{M}_{GS}^*) = GED(P', Q')$ . Let  $\mathcal{M}_{AGS}$  be the matching returned by  $AGED(P', Q', (4\sqrt{2} + 6)g)$ . We have the following lemma.

**Lemma 3.** *If  $GED(P, Q) \leq g$ , with a probability at least  $1 - \frac{1}{n^c}$ , at least one of the  $\lceil c \lg n \rceil$  iterations will return a matching  $\mathcal{M}_{AGS}$ .*

**Proof:** Recall, we said pairs of points are *close* if their distance is less than or equal to  $\Delta$ . Similar to (2), and using Observation 1, we have

$$\begin{aligned} C_{GS}(\mathcal{M}_G^*) &= |UM_{\mathcal{M}_G^*}| + 0 \cdot (|OC_{\mathcal{M}_G^*}| + |OF_{\mathcal{M}_G^*}|) \\ &\quad + \sum_{(i,j) \in DC_{\mathcal{M}_G^*}} \text{dist}(p'_i, q'_j) + \sum_{(i,j) \in DF_{\mathcal{M}_G^*}} \text{dist}(p'_i, q'_j) \\ &\leq |UM_{\mathcal{M}_G^*}| + \Delta \cdot |DC_{\mathcal{M}_G^*}| + \sum_{(i,j) \in DF_{\mathcal{M}_G^*}} \left( \text{dist}(p_i, q_j) + 2\sqrt{2}\Delta \right). \\ &= |UM_{\mathcal{M}_G^*}| + \sum_{(i,j) \in DF_{\mathcal{M}_G^*}} \text{dist}(p_i, q_j) + \Delta |DC_{\mathcal{M}_G^*}| + 2\sqrt{2}\Delta |DF_{\mathcal{M}_G^*}| \end{aligned}$$

If  $C_G(\mathcal{M}_G^*) \leq g$ , then

$$C_{GS}(\mathcal{M}_{GS}^*) \leq g + \Delta |DC_{\mathcal{M}_G^*}| + 2\sqrt{2}\Delta |DF_{\mathcal{M}_G^*}|.$$

We have the same observation for  $DF_{\mathcal{M}_G^*}$  as before, that is there are at most  $\frac{g}{\Delta}$  pairs in  $DF_{\mathcal{M}_G^*}$ . Using the same algebra as (4), we have  $\mathbb{E}(|DC_{\mathcal{M}_G^*}|) \leq \frac{2g}{\Delta}$ . So,

$$\mathbb{E}(C_{GS}(\mathcal{M}_{GS}^*)) \leq g + \Delta \cdot \frac{g}{\Delta} + 2\sqrt{2}\Delta \frac{2g}{\Delta} = 2\sqrt{2}g + 3g.$$

According to Markov's inequality, we know

$$P\left(C_{GS}(\mathcal{M}_{GS}^*) \geq (4\sqrt{2} + 6)g\right) \leq \frac{1}{2}.$$

In Section 4.2, we prove that if  $C_{GS}(\mathcal{M}_{GS}^*) = GED(P', Q') \leq (4\sqrt{2} + 6)g$ , then  $AGED(P', Q', (4\sqrt{2} + 6)g)$  will return a constant approximate matching  $\mathcal{M}_{AGS}$ . So, if we test  $AGED(P', Q', (4\sqrt{2} + 6)g)$   $\lceil c \lg n \rceil$  times (using different grids each time), with a probability at least  $1 - \frac{1}{n^c}$ , at least one  $AGED(P', Q', (4\sqrt{2} + 6)g)$  will return a matching  $\mathcal{M}_{AGS}$ . We conclude the proof of Lemma 3.  $\square$

Finally, from Observation 1, for every pair  $(i, j)$  in  $\mathcal{M}_{AGS}$ , we have  $dist(p_i, q_j) \leq dist(p'_i, q'_j) + 2\sqrt{2}\Delta$ . We can now return points to their original positions:

$$\begin{aligned}
C_G(\mathcal{M}_{AGS}) &= |UM_{\mathcal{M}_{AGS}}| + \sum_{(i,j) \in DC_{\mathcal{M}_{AGS}}} dist(p_i, q_j) + \sum_{(i,j) \in DF_{\mathcal{M}_{AGS}}} dist(p_i, q_j) \\
&\quad + \sum_{(i,j) \in OC_{\mathcal{M}_{AGS}}} dist(p_i, q_j) + \sum_{(i,j) \in OF_{\mathcal{M}_{AGS}}} dist(p_i, q_j) \\
&\leq |UM_{\mathcal{M}_{AGS}}| + \sum_{(i,j) \in DC_{\mathcal{M}_{AGS}}} dist(p'_i, q'_j) + \sum_{(i,j) \in DF_{\mathcal{M}_{AGS}}} dist(p'_i, q'_j) + \sum_{(i,j) \in OC_{\mathcal{M}_{AGS}}} dist(p'_i, q'_j) \\
&\quad + \sum_{(i,j) \in OF_{\mathcal{M}_{AGS}}} dist(p'_i, q'_j) + 2\sqrt{2}\Delta (|DC_{\mathcal{M}_{AGS}}| + |DF_{\mathcal{M}_{AGS}}| + |OC_{\mathcal{M}_{AGS}}| + |OF_{\mathcal{M}_{AGS}}|) \\
&\leq O(1) \cdot (4\sqrt{2} + 6)g + n \cdot 2\sqrt{2}\Delta.
\end{aligned}$$

Recall,  $\Delta = \frac{g\alpha}{n}$ . If we obtain a matching  $\mathcal{M}_{AGS}$  during an iteration where  $g \geq C_G(M_G^*) = GED(P, Q) \geq \frac{1}{2}g$ , then  $C_G(\mathcal{M}_{AGS}) \leq O(g\alpha) = O(\alpha) \cdot GED(P, Q)$ . Finishing with the same argument as in Theorem 1, we conclude our proof of Theorem 2.

## 4 Constant Approximation Algorithm $AGED(P', Q', k)$

Recall that our constant factor approximation algorithm for GED of grid corner points is based on a known  $O(n + k^2)$  time exact algorithm for string edit distance [23]. We first describe this exact algorithm for strings, which we refer as  $SED(S, T, k)$ , in Section 4.1. Then in Section 4.2, we modify this string algorithm to obtain an  $O(1)$ -approximate matching for edit distance between point sequences  $P'$  and  $Q'$  assuming the points lie on the corners of grid cells and  $GED(P', Q') \leq k$ .

### 4.1 The exact $O(n + k^2)$ string edit distance algorithm

**Dynamic programming matrix and its properties.** Let  $S = \langle s_1, s_2, \dots, s_n \rangle$  and  $T = \langle t_1, t_2, \dots, t_n \rangle$  be two strings of length  $n$ . Let  $D$  be the  $(n + 1) \times (n + 1)$  matrix where  $D(i, j)$  is the edit distance between substrings  $S_i = \langle s_1, s_2, \dots, s_i \rangle$  and  $T_j = \langle t_1, t_2, \dots, t_j \rangle$ . We give a label  $h$  to every diagonal in this matrix such that for any entry  $(i, j)$  in this diagonal,  $j = i + h$ . See Fig. 1 (a).

Recall, we aim to minimize only the number of insertions and deletions to turn  $S$  into  $T$ . There are four important properties in this matrix which are used in the  $O(n + k^2)$  time algorithm.

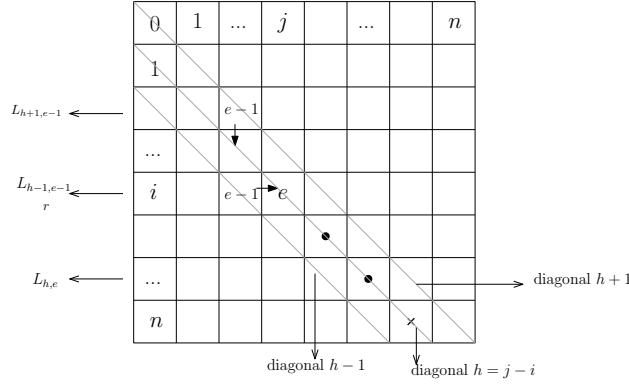
$$\textbf{Property 1. } D(i, j) = \min \begin{cases} D(i - 1, j) + 1 \\ D(i, j - 1) + 1 \\ D(i - 1, j - 1) + |s_i t_j| \end{cases} \quad \text{where } |s_i t_j| = \begin{cases} 0, & \text{if } s_i = t_j \\ \infty, & \text{otherwise} \end{cases}.$$

**Property 2.**  $D(i, 0) = i$ , and  $D(0, j) = j$ .

**Property 3.**  $D(i, i + h)$  is even if and only if  $h$  is even.

**Property 4.**  $D(i, j) - D(i - 1, j - 1) \in \{0, 2\}$ .

Property 4 can be easily derived from Property 3 and induction on  $i + j$  (see Lemma 3 of [10]). From Property 4, we know all the diagonals are non-decreasing. In particular, all values on diagonal  $h$  are greater than  $|h|$  considering Property 2. So, we can just search the band from diagonal  $-k$  to  $k$  if the edit distance between  $S$  and  $T$  is at most  $k$ .



**Figure 1.** (a) The diagonal containing any entry  $(i, i + h)$  is diagonal  $h$  (b) The algorithm slides down the diagonal until finding an entry representing distinct characters. A circle means the corresponding two characters are the same; a cross means they are different

**Algorithm for edit distance at most  $k$ .** We use a greedy approach to fill the entries along each diagonal. For each value  $e \in \{0, \dots, k\}$  (the outer loop), we locate the elements whose value is  $e$  by inspecting diagonals  $-e$  to  $e$  (the inner loop). Finally, we return the best matching if  $D(n, n)$  is covered by the above search. Otherwise, the edit distance is greater than  $k$ .

The key insight is that we can implicitly find all entries containing  $e$  efficiently in each round. We first define  $L_{h,e}$  as the row index of the *farthest*  $e$  entry in diagonal  $h$ .

**Definition 2.**  $L_{h,e} = \max\{i \mid D(i, i + h) = e\}$ .

Note by Property 3,  $L_{h,e}$  is well-defined only if  $h \equiv e \pmod{2}$ . Observe that all values on diagonal  $h$  are at least  $|h|$ , so we can define our initial values as:

$$L_{h,|h|-2} = \begin{cases} |h| - 1, & \text{if } h < 0; \\ -1, & \text{otherwise} \end{cases}, \text{ where } h \in [-k, k].$$

Let  $r = \max\{L_{h-1,e-1}, L_{h+1,e-1} + 1\}$ . Then,  $D(r, r + h) = e$  by Properties 1 and 4. Also, if  $D(r, r + h) = e$  and  $s_{r+1} = t_{r+1+h}$ , then  $D(r + 1, r + 1 + h) = e$ . From these observations, we can compute  $L_{h,e}$  in each inner loop using Algorithm 3 below.

---

**Algorithm 3:** Computing  $L_{h,e}$  in each inner loop

---

```

1  $r := \max\{L_{h-1,e-1}, L_{h+1,e-1} + 1\}$ 
2 while  $r + 1 \leq n$ ,  $r + h + 1 \leq n$ , and  $s_{r+1} == t_{r+1+h}$  do
3   |  $r := r + 1$ ; /* slide */
4 end
5 if  $r > n$  or  $r + h > n$  then
6   |  $L_{h,e} := \infty$ 
7 else
8   |  $L_{h,e} := r$ 
9 end

```

---

We call lines 2 through 4 “the slide”. It is straightforward to recover the optimal matching by using the  $L_{h,e}$  values to trace backwards through the dynamic programming matrix. Fig. 1 (b) demonstrates this process.

We can perform slides in constant time each after some  $O(n)$ -time preprocessing at the beginning of the algorithm. Specifically, the length of a slide can be computed using a single lowest common ancestor query in the suffix tree of a string based on  $S$  and  $T$  [23]. The overall running time is  $O(n + k^2)$ .

## 4.2 $O(1)$ -approximation algorithm for GED of grid-snapped points

**Notations.** Similar to the string algorithm, we have a dynamic programming matrix;  $D'(i, j)$  is the edit distance between subsequences  $P'_i = \langle p'_1, \dots, p'_i \rangle$  and  $Q'_j = \langle q'_1, \dots, q'_j \rangle$  of points at the corners of grid cells. This matrix also meets Property 1 stated earlier except that we use  $\text{dist}(p'_i, q'_j)$  instead of  $|s_i t_j|$ . In addition, we also have the following property which is a refinement of Property 4.

**Property 5.**  $D'(i, j) - D'(i - 1, j - 1) \in [0, 2]$ .

Clearly, the upper bound is 2 (just unmatched  $p_i$  and  $q_j$ ). The lower bound can be proved by induction. Because the values in any diagonal are non-decreasing, we need only consider diagonals  $-k$  through  $k$ .

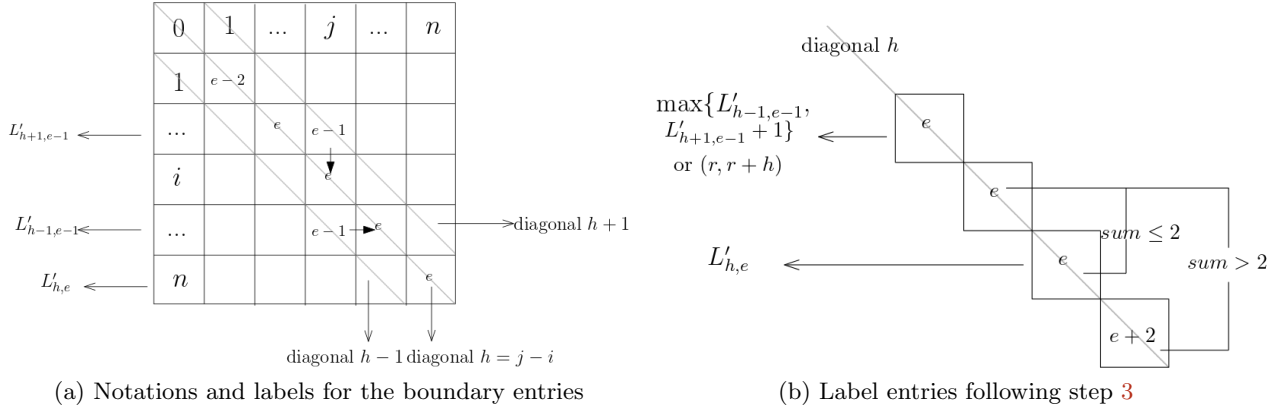
**(Implicit) label rules.** To obtain an approximate matching for the edit distance of snapped point sequences, we now label each entry in the dynamic programming matrix with an approximately tight lower bound on its value. Inspired by the string algorithm, we use non-negative integers for our labels, and the entries of any diagonal  $h$  only receive labels  $e$  where  $e \equiv h \pmod{2}$ . Let  $LA(i, j)$  be the label of entry  $(i, j)$  and  $L'_{h,e}$  be the row index of the farthest entry whose label is  $e$  in diagonal  $h$ .

**Definition 3.**  $L'_{h,e} := \max\{i \mid LA(i, i + h) = e\}$ .

For each  $e$  from 0 to  $k$ , for each diagonal  $h$  where  $h \equiv e \pmod{2}$ , we (implicitly) assign label  $e$  to some entries on diagonal  $h$ .

1. If  $h = -e$  or  $e$ , i.e., this is the first iteration to assign labels to this diagonal, then we label the very beginning entry in diagonal  $h$  as  $e$ , i.e., if  $h = -e$ ,  $LA(-h, 0) = e$ ; otherwise,  $LA(0, h) = e$ .
2. We define a *start entry*  $(r, r + h)$  for each diagonal  $h$ . If  $h = -e$  or  $e$ ,  $r$  is the row index of the first one entry in diagonal  $h$  ( $r = |h|$  or 0); otherwise,  $r = \max\{L'_{h-1, e-1}, L'_{h+1, e-1} + 1\}$ .
3. We assign the label  $e$  to entries  $(r, r + h)$  to  $(r + s, r + h + s)$  where  $\sum_{i=r+1}^s \text{dist}(p'_i, q'_{i+h}) \leq 2$  and  $\sum_{i=r+1}^{s+1} \text{dist}(p'_i, q'_{i+h}) > 2$ .  $L'_{h,e} := r + s$ . These entries correspond to a slide in the string algorithm.
4. Finally, if  $(r - 1, r + h - 1)$  is unlabeled, we go backward up the diagonal labeling entries as  $e$  until we meet an entry that has been assigned a label previously. (Again, this step is implicit. As explained below, the actual algorithm only finds the  $L'_{h,e}$  entries.)

Fig. 2 illustrates our rules.



**Figure 2.** Notations and rules for approximating SGED

**Computing an approximately optimal matching.** Assume we have set the initial values. Our algorithm only needs to compute each  $L'_{h,e}$  as before. See Algorithm 4. Then, we guarantee the following theorem:

**Theorem 3.** Suppose  $GED(P', Q') \leq k$ . We can recover a matching  $M_{GS}^{*'}$  using the values  $L'_{h,e}$  from Algorithm 4. The cost of  $M_{GS}^{*'}$  for point sequences  $P', Q'$  is less and equal to  $3GED(P', Q')$ .

In short, we argue each label  $LA(i, j) \leq D'(i, j)$ . We then follow a path through the matrix as suggested by the way we pick labels in Algorithm 4. The final matching has cost at most  $3LA(n, n)$  which is less and equal to  $3GED(P', Q')$ . The full proof appears in Section 4.3

---

**Algorithm 4:** Computing  $L'_{h,e}$  for the fixed  $h$  and  $e$

---

```

1  $r := \max\{(L'_{h-1, e-1}), (L'_{h+1, e-1} + 1)\}$ 
2  $sum := 0$ 
3 while  $r + 1 \leq n, r + h + 1 \leq n$ , and  $(sum + dist(p'_{r+1}, q'_{r+h+1}) \leq 2)$  do
4    $r := r + 1$ 
5    $sum := sum + dist(p'_r, q'_{r+h})$ 
6 end
7 if  $r > n$  or  $r + h > n$  then
8    $L'_{h,e} := \infty$ 
9 else
10   $L'_{h,e} := r$ 
11 end

```

---

We conclude by discussing the time complexity for our algorithm. Using the same  $O(n)$  pre-processing as in [23], we can slide down maximal sequences of consecutive entries  $(r, r + h)$  with  $dist(p'_r, q'_{r+h}) = 0$  in constant time per slide. Let  $\Delta$  be the cell side length of the grid whose cell corners contain points of  $P'$  and  $Q'$ . For  $dist(p'_r, q'_{r+h}) \neq 0$ , we know  $dist(p'_r, q'_{r+h}) \geq \Delta$  from Observations 1. Therefore, we only need to manually add distances and restart faster portions of each slide of distances summing to 2 a total of  $\frac{2}{\Delta}$  times. Thus, the total running time is

$$O(n + \sum_{e=0}^k \sum_{h=-e}^e \frac{1}{\Delta}) = O(n + \frac{k^2}{\Delta}).$$

### 4.3 Proof of Theorem 3

We have the following properties for our labels along with the following lemma.

**Property 6.**  $LA(i, i + h) - LA(i + 1, i + 1 + h) \in \{0, 2\}$ .

**Property 7.**  $LA(i, i + h) - LA(i - 1, i + h) \in \{-1, 1\}$  and  $LA(i, i + h) - LA(i, i + h - 1) \in \{-1, 1\}$ .

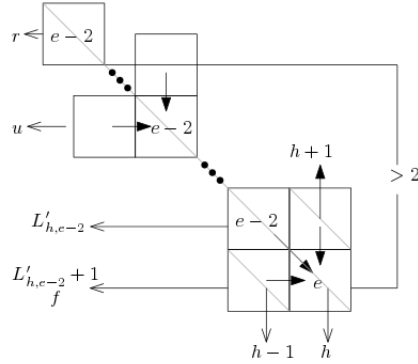
**Lemma 4.** For every entry  $(i, j)$  assigned a label,  $LA(i, j) \leq D'(i, j)$ .

Note that in particular,  $LA(n, n) \leq GED(P', Q')$ .

**Proof:** From Property 5, we only need to prove  $e$  is the lower bound of the first entry whose label is  $e$  in each diagonal  $h$ .

We proceed by induction on  $e$ .

1. If  $e = 0$ , we only label the first entry in diagonal 0 as 0. We have  $0 \leq D'(0, 0) = 0$ . If  $e = 1$ , then for diagonals 1 and  $-1$ , we have  $1 \leq D'(0, 1) = D'(1, 0) = 1$ .
2. Assume Lemma 4 for labels less than  $e$ . For  $e$ , we consider the diagonals  $h = -e$  to  $e$ : If  $h = -e$  or  $e$ , we know  $e \leq D'(|h|, 0) = e$  or  $e \leq D'(0, h) = e$ . Otherwise, let  $(f, f + h)$  be the first entry in diagonal  $h$  whose label is  $e$ . From Property 6,  $f = L'_{h, e-2} + 1$ . Fig. 3 shows the notations. From the refined Property 1, we need to discuss



**Figure 3.** We compute the lower bound of entries which are labeled as  $e$

three cases:

(a)  $D'(f, f + h) = D'(f - 1, f + h) + 1$ .

From Property 7, we know  $LA(f - 1, f + h) = e - 1$  or  $e + 1$ .

- If  $LA(f - 1, f + h) = e - 1$ ,  $D'(f - 1, f + h) \geq e - 1$  from our assumption. So,  $D'(f, f + h) = D'(f - 1, f + h) + 1 \geq e - 1 + 1 = e$ .
- If  $LA(f - 1, f + h) = e + 1$ , then we know  $L'_{h+1, e-1}$  is less than  $f - 1$ . From our assumption and non-decreasing property,  $e - 1 \leq D'(L'_{h+1, e-1}, L'_{h+1, e-1} + h + 1) \leq D'(f - 1, f + h - 1)$ . So,  $D'(f, f + h) \geq D'(f - 1, f + h - 1) + 1 \geq e$ .

(b)  $D'(f, f + h) = D'(f, f + h - 1) + 1$ .

This case is symmetric to the one above.

- (c)  $D'(f, f+h) = D'(f-1, f+h-1) + \text{dist}(p'_f, q'_{f+h})$ .  
 $LA(f-1, f+h-1) = e-2$ , because  $f-1 = L'_{h,e-2}$ . Let  $r$  be the row index of the first entry on the slide with label  $e-2$  in diagonal  $h$ , i.e.,  $r = \max\{L'_{h-1,e-3}, L'_{h+1,e-3} + 1\}$ . See Fig. 3. We define  $u$  as the row index of the first entry walking backward from entry  $(f, f+h)$  along the diagonal  $h$  where  $D'(u, u+h) = \min\{D'(u, u+h-1), D'(u-1, u+h)\} + 1$ .

- If  $u > r$ , like Fig. 3, then  $u > L'_{h-1,e-3}$  and  $u-1 > L'_{h+1,e-3}$ . We have

$$D'(u, u+h-1) \geq D'(L'_{h-1,e-3} + 1, L'_{h-1,e-3} + h) \geq e-1$$

and

$$D'(u-1, u+h) \geq D'(L'_{h+1,e-3} + 1, L'_{h+1,e-3} + h + 2) \geq e-1.$$

Therefore,

$$D'(u, u+h) = \min\{D'(u, u+h-1), D'(u-1, u+h)\} + 1 \geq e.$$

By Property 5, we have  $D'(f, f+h) \geq e$  as well.

- If  $u \leq r$ , then

$$\begin{aligned} D'(f, f+h) &= D'(r, r+h) + \sum_{i=r+1}^f \text{dist}(p'_i, q'_{i+h}) \\ &> e-2 + 2 = e. \end{aligned}$$

Having considered all the cases, we conclude the proof of Lemma 4. □

**The bounds for the approximate matching  $C_{GS}(\mathcal{M}_{AGS})$ .** From Algorithm 4, we note the label increases correspond to not matching a point in Line 1, and slides correspond to matching points. Let  $\mathcal{M}_{AGS}$  be the resulting matching. So,

$$\begin{aligned} C_{GS}(\mathcal{M}_{AGS}) &= |UM_{\mathcal{M}_{AGS}}| + \sum_{(i,j) \in \mathcal{M}_{AGS}} \text{dist}(p'_i, q'_j) \\ &\leq LA(n, n) + 2 \cdot LA(n, n) \leq 3LA(n, n) \leq 3GED(P', Q'). \end{aligned}$$

We conclude the proof of Theorem 3.

**Acknowledgements** The authors would like to thank Anne Driemel and Benjamin Raichel for helpful discussions.

## References

- [1] K. Fox, X. Li, in *Proceedings of the 30th International Symposium on Algorithms and Computation* (2019), pp. 26:1–26:16
- [2] P.K. Agarwal, K. Fox, J. Pan, R. Ying, in *Proceedings of the 32nd International Symposium on Computational Geometry* (2016), pp. 6:1–6:16
- [3] A. Stojmirovic, Y.k. Yu, *Journal of Computational Biology* **16**(4), 579 (2009)

- [4] L. Chen, R. Ng, in *Proceedings of the 30th International Conference on Very Large Databases* (2004), pp. 792–803
- [5] L. Chen, M.T. Özsu, V. Oria, in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (2005), pp. 491–502
- [6] P.F. Marteau, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 306 (2009)
- [7] S. Sankararaman, P.K. Agarwal, T. Mølhave, J. Pan, A.P. Boedihardjo, in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2013), pp. 234–243
- [8] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, E. Keogh, *Data Mining and Knowledge Discovery* **26**(2), 275 (2013)
- [9] R.A. Wagner, M.J. Fischer, *Journal of the ACM* **21**(1), 168 (1974)
- [10] E. Ukkonen, *Information and Control* **64**(1-3), 100 (1985)
- [11] W.J. Masek, M.S. Paterson, *Journal of Computer and System Sciences* **20**(1), 18 (1980)
- [12] O. Gold, M. Sharir, *ACM Transactions on Algorithms* **14**(4), 50 (2018)
- [13] R. Impagliazzo, R. Paturi, *J. Comp. Sys. Sci.* **62**(2), 367 (2001)
- [14] K. Bringmann, in *Proceedings of the IEEE 55th Annual Symposium on Foundations of Computer Science* (2014), pp. 661–670
- [15] A. Abboud, A. Backurs, V.V. Williams, in *Proceedings of the IEEE 56th Annual Symposium on Foundations of Computer Science* (2015), pp. 59–78
- [16] K. Bringmann, M. Künnemann, in *Proceedings of the IEEE 56th Annual Symposium on Foundations of Computer Science* (2015), pp. 79–97
- [17] A. Backurs, P. Indyk, in *Proceedings of the 47th Annual ACM Symposium on Theory of Computing* (2015), pp. 51–58
- [18] K. Bringmann, W. Mulzer, *JoCG* **7**(2), 46 (2016)
- [19] T.M. Chan, Z. Rahmati, *Information Processing Letters* pp. 72–74 (2018)
- [20] W. Kuszmaul, in *Proceedings of the 46th International Colloquium on Automata, Languages and Programming* (2019)
- [21] A. Andoni, R. Krauthgamer, K. Onak, in *Proceedings of the IEEE 51st Annual Symposium on Foundations of Computer Science* (2010), pp. 377–386
- [22] D. Chakraborty, D. Das, E. Goldenberg, M. Koucky, M. Saks, in *Proceedings of the 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2018), pp. 979–990
- [23] G.M. Landau, E.W. Myers, J.P. Schmidt, *SIAM Journal on Computing* **27**(2), 557 (1998)



- [24] A. Andoni, N.S. Nosatzki, in *Proceedings of the IEEE 61st Annual Symposium on Foundations of Computer Science* (2020). To appear.
- [25] S. Har-Peled, *Geometric approximation algorithms* (American Mathematical Soc., 2011), chap. 11, Random Partition via Shifting