

Time Range Queries for Hereditary Properties

6.854 Final Project

Arsen Mamikonyan, Hayk Saribekyan

December 13, 2016

Abstract

Time range queries are important for analysing large datasets with timestamps. Such datasets are common in computational geometry, where a sequence of points with timestamps often corresponds to a trajectory of an object. Time range queries in this case check a certain property of section of the trajectory.

In this paper we review recent general frameworks that can be used to handle property testing queries for time ranges, and discuss their implementations for solving certain problems in computational geometry. The results are limited to *hereditary* properties, which cover a large variety of interesting problems.

For two different hereditary properties, we compared the performance of efficient an algorithm and a naive one in practice. We observed that the efficient algorithms require more than ten times the amount of code in certain cases, but the large constant factor associated with it is compensated when working with large datasets.

1 Introduction

TODO: add normal references (maybe to envelopes), see the first paper, where in the first page there are a lot of motivational references.

With the abundance of GPS and other movement tracking sensors there is a large amount of timestamped location data. The movement of an object in space can be modelled by a sequence of points $S = s_1, \dots, s_n$. To analyse the trajectory of the movement, one may want to check a certain property P for portions $S[i, j] = s_i, \dots, s_j$ of S for given values of i and j . This paper discusses how to efficiently address such queries when P satisfies the following two properties:

- P is boolean.
- P is hereditary. This means that for a given sequence S , if $P(S)$ is true, then $P(S')$ is true for any continuous subsequence S' of S .

In this paper we discuss and implement two properties for a sequence of points $S = s_1, \dots, s_n$:

- *Monotonicity*: S is *monotone* if there is a direction vector v such that **TODO**. In terms of trajectories, monotonicity shows whether the object travelled more or less in the same direction.
- *Closeness*: S is *close* if any pair of points in S is at most of distance 1. One could use closeness queries to detect if a moving object mostly stayed in the same surrounding.

For a property P satisfying the restrictions above, let $j^*(i)$ be the largest index j such that P holds for $S[i, j^*(i)] = s_i, \dots, s_{j^*(i)}$. Now notice, that since P is a hereditary property $P(S[i, j]) = \text{true}$ if and only if $i \leq j \leq j^*(i)$. Therefore, if we construct $j^*(i)$ we can answer property testing queries in $O(1)$. Therefore, for the rest of the paper our goal will be to efficiently compute $j^*(i)$.

Bokal et al. [1] propose an algorithms that compute $j^*(i)$ in $O(n)$ time for monotonicity, and in $O(n \log^2 n)$ time for closeness. Chan and Pratt [2] describe a different algorithm to achieve the same result for closeness, but also improve it to $O(n \log n)$ using fractional cascading.

The rest of the paper is organized as follows: Section 2 describes the naive algorithms we developed to solve the problems we selected; in Section 3 we review the general framework that is set in [1] for solving time range query problems, and show how it is applied to compute $j^*(i)$ for monotonicity. Section 4 reviews the framework by [2] and its application for the closeness property. We then give some implementation details in Section 5, and describe the experimental setup and our tests in Section 6.

2 Naive Algorithms

Suppose that for a sequence of length n a property P can be checked in $T_P(n)$ time. Then, one can compute $j^*(i)$ using binary search for each i . On a step of a binary search concerning a range $[l, r]$ the algorithm would spend $T_P(r - l)$ time. The total runtime of this simple algorithm is $O(nT_P(n) \log n)$. Call this algorithm *SN* (stands for super-naive). For the problems of monotonicity and closeness we can do better.

2.1 Monotonicity

To detect whether a sequence is monotone or no, we can process the points while keeping a set of polar angles (field-of-view or FoV) that a direction vector v can have. Notice that the FoV is always an interval. When a new point arrives we can update the FoV in constant time. If the FoV ever becomes empty, then the given sequence is not monotone. This algorithm runs in $O(n)$ time for a sequence of length n . Using this submodule for *SN* we would get a $O(n^2 \log n)$ algorithm for computing $j^*(i)$ for all values of i .

However, doing a binary search for monotonicity is redundant, since we will do the same computation many times. Instead, for each i proceed with the

FoV computation until it is empty, which will happen precisely when we reach $j^*(i) + 1$. The resulting algorithm runs in $O(n^2)$ time.

2.2 Closeness

The simple method of closeness-check takes $O(n^2)$ for a sequence of n points, because one has to check all pairs of points. This results in SN runtime of $O(n^3 \log n)$, which is super-slow. We can improve the closeness-check to $O(n \log n)$ using furthest-point Voronoi diagrams, but that would not qualify for SN and would also take $O(n^2 \log^2 n)$ time to compute, which is also not fast.

Define $k^*(i)$ as the largest index such that $d(s_i, s_j) \leq 1$ for all $j \leq k^*(i)$. Clearly, $k^*(i) \geq j^*(i)$ because we are checking for a weaker property. Also, notice that k^* can be computed in $O(n^2)$ for all points by calculating all pairwise distances.

Noticing that $j^*(i) = \min(j^*(i+1), k^*(i))$, we can calculate $j^*(i)$ for all points in $O(n)$ extra time if we are given all values for $k^*(i)$. So we found a $O(n^2)$ algorithm for computing closeness property for all subsequences.

3 Monotonicity in Linear Time

Bokal et al. introduced an elegant framework to deal with range queries for hereditary problems. The key idea in their work is to greedily split the given set of points into ranges, defined by anchor points, and solve for each using a divide and conquer. To simplify the reasoning we first define a property matrix A_P such that

$$A_P(i, j) = \begin{cases} 1 & P(S[i, j]) = \text{true} \\ 0 & P(S[i, j]) = \text{false} \end{cases}$$

See Figure 3 for an example. Suppose that there is an algorithm J_P that, for a given i , finds $j^*(i)$ (notice that even if such J_P exists, we are not happy to run it n times for each i). We use J_P to find *anchor* points in S as follows. Let a_k be the index of k^{th} anchor point in S . We define $a_1 = s_1$, $a_k = \max(1 + a_{k-1}, j^*(a_{k-1}))$ for $k > 1$. Thus, the anchor points can be found using J_P . Figure 3 shows the anchor points on A (leftmost image).

Also shown in Figure 3 are rectangles formed using the anchor points. If we solve a rectangle (find the curve in A separating 1's from 0's) for each rectangle, then we will be done, as stated in Lemma 2.1 from Bokal et al.

Lemma 2.1 Assume that we have the following two subroutines for sequence S and a property P :

- (a) For any $a = 1, \dots, n$ find the $j^*(a)$, taking $T_{greedy}(j^*(a) - a)$
- (b) Solve a rectangle with anchor points at opposite vertices (anchored rectangles). Suppose this module takes $T_{rect}(\text{height}(R) + \text{width}(R))$.

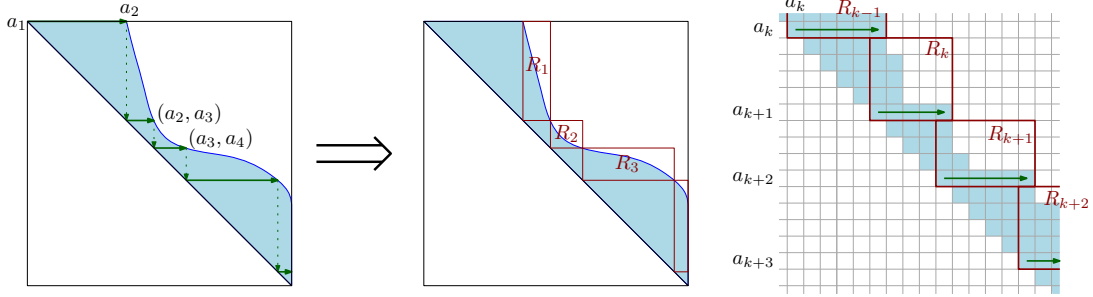


Figure 1: The property matrix A_P . The blue regions show the locations that are 1. The left image illustrates the anchor points. The middle one shows the greedily constructed rectangles from the anchor points. On the right side is the detailed view of the property matrix. Figure from [1].

Both of these modules are property specific, but Bokal et al. make another interesting generic step that explains how solving rectangles can be done using a divide-and-conquer approach. We will, however, skip that step and focus on the problem of monotonicity. It turns out that monotonicity does not require the divide-and-conquer approach to solve anchored rectangles.

3.1 Solving Monotonicity

Section 2.1 discusses how to find $j^*(i)$ in linear time for a given i by using intervals of polar angles. This algorithm will serve as the subroutine for Lemma 2.1(a).

Consider an anchored rectangle R with a lower-left corner (a, a) , width w and height h .

- By traversing the points $S[a, a+w]$, we can compute for each $j = a, \dots, a+w$ the interval of polar angles I_j such that $S[a, j]$ is monotone with respect to all directions in I_j .
- By traversing the points $S[a-h, a]$ in reverse order, a similar set I_i can be computed for $a-h \leq i < a$, representing the set of monotone directions for $S[i, a]$.

If $i < a < j$, then $S[i, j]$ is monotone if and only if I_i and I_j have a non-empty intersection. Thus, for each $i = a-h, \dots, a$, $j^*(i)$ is the largest index such that $I_i \cap I_{j^*(i)} \neq \emptyset$. Noting that search for $j^*(i+1)$ can start from $j^*(i)$, we can compute $j^*(i)$ for all $i \in [a-h, a]$ in linear time.

These two subroutines, together with Lemma 2.1 give a linear solution to the monotonicity problems. We have evaluated Naive algorithm (from Section 2) and this algorithm in Section 6.

4 something and Closeness TODO

TODO: Arsen, this section should describe some preliminaries of the algorithm (e.g. the notion of polyarcs), so that there is coherency between the sections.

The novelty of the algorithm from Chan, Pratt is how they combine 1D range tree, with secondary structure that stores circle intersections. While the algorithm is elegant, its implementation is not. In particular, it is necessary to implement an intersection algorithm for arbitrary *polyarcs*. A polyarc is a convex object, similar to a polygon, but instead of straight edges each side is an arc of a circle. The algorithm for the intersection of polyarcs is similar to that of polygons, however there are many more edge cases to consider.

5 Implementation Details

In the papers reviewed here ([1, 2]) there are at least 6 algorithms presented. We wanted to implement algorithm that use frameworks of both papers. In order to limit the size of the implementation we decided to pick monotonicity property for its simplicity (and elegance), which was solved in Bokal et al. [1]. We chose to implement the closeness testing algorithm from [2] because it used many ideas from the class, for example the intersection of convex envelopes in 2 dimensions, range trees.

The implementations of naive algorithms as well as of the efficient monotonicity testing algorithm are relatively easy, so in this paper we skip discussing them. We implemented the $O(n \log^2 n)$ algorithm for solving closeness testing from [2].

The algorithm has two distinct parts: a high-level tree-range data structure, and a lower-level data structure for storing *polyarcs* (see Section 4). A polyarc is a convex geometric shape, which is defined by a clockwise set of its vertices like a polygon. Unlike a polygon, each edge of a polyarc is an arc of a circle. The algorithm needs to be able to find the intersection of two polyarcs, and also decide whether a query point is inside the polyarc.

Both of these algorithms are similar to those of polygons, but the majority of the implementation effort was spent to handle the cases specific to polyarcs. Our representation of polyarcs has two “sides”: upper and lower, respectively the upper part of the complex shape and the lower one. To intersect two polyarcs we first found a minimal upper envelope i.e. for each x coordinate found recorded the polyarc, which had a lower upper side. Similarly, we computed a maximal lower envelope. The area above the maximal lower envelope and minimal upper envelope is the intersection of two polyarcs. Each of these steps can be performed in linear time by a sweep line technique.

The envelope representation of polyarcs is also helpful for querying whether a point is inside the polyarc or no. One needs to simply do a binary search for the x coordinate of the candidate point. If a vertical line at that location crosses the higher envelope on a larger value of y than the lower envelope, then the query point is inside the polyarc.

Overall, we wrote more than 1200 lines of code in C++ for the implementation of the algorithms, and 800 lines of code for testing. Large part of the code is dedicated to the closeness algorithm in $O(n \log^2 n)$. We used C++ for the implementation so that the overhead that the language adds to the runtime is minimized. A simple language to write code in is Python, but we think that in this case the strong type checking of C++ helped us, because of a complex structure of the code.

6 Experimental Results

- mention that we have used random walk to mimic reality. in that case the good algorithm for monotonicity is not that good. but for the worst case, it's great.

6.1 Monotonicity

TODO: Describe Naive algorithm which is a $O(n^2)$ algorithm.

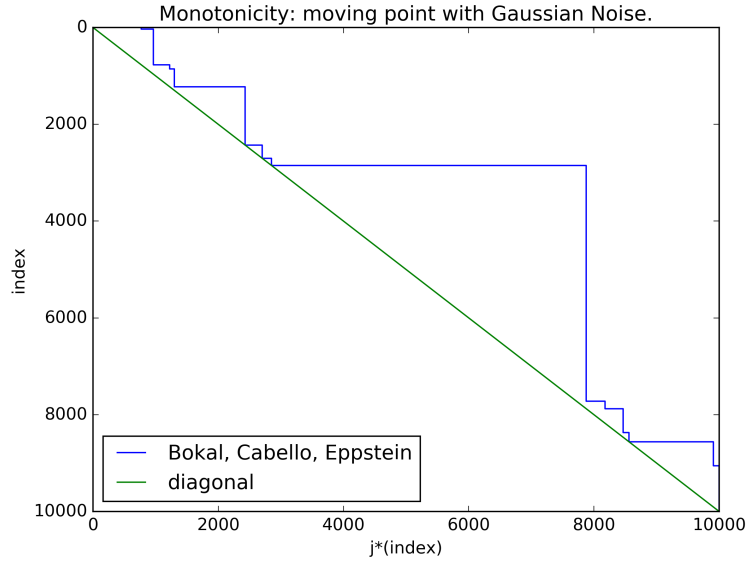


Figure 2: Monotonicity algorithm results for moving point with Gaussian noise.

First let's look at the monotonicity results. Here we use the following dataset. We have a point moving along x axis in the positive direction with speed 1 (i.e. 1 per index), we add Gaussian noise to this point to make the problem interesting. In Figure 2 you can see results for noise with standard deviations

$\sigma_y = 0.05, \sigma_x = 1$. After we generate the dataset, we run Bokal et. al algorithm on the dataset and the answer boundaries of the matrix A that indicates if for points in range $[i, j]$ exists a common direction that has positive dot product with each of the displacements.

Figure 3 shows runtime differences in milliseconds between Naive algorithm and algorithm we presented from Bokal et al.

Bokal et al	Naive
0.616	3.43
2.008	12.799
6.038	45.436
19.33	152.641
61.926	428.032
194.457	1454.634

Figure 3: Runtimes (in milliseconds) of Naive and Bokal et. al

6.2 Diameter

Now the dataset is a random walk starting at the origin. At each step we uniform randomly pick a direction, and move 0.3 distance in that direction. We keep doing this until we have enough points for an experiment.

TODO: Describe Naive algorithm which is a $O(n^2)$ algorithm.

Figure 4 shows boundaries of points that satisfy all points in range $[i, j]$ fall into some circle with radius 1.

Figure 5 will show results comparing runtimes of Naive algorithm with algorithm we presented from Bokal et al.

7 Conclusion

It was really challenging to do an implementation project in computational geometry, some sentences in the papers take 100+ lines of code to implement. Thus, for small datasets, added complexity is not worth the speedup gained by the algorithm.

This was a great project otherwise, we've implemented couple of algorithms that we have seen during the lecture - sweep line, range tree, intersection of convex polygons using envelopes. Also couple that we had not found in Bokal et al and Chan, Pratt.

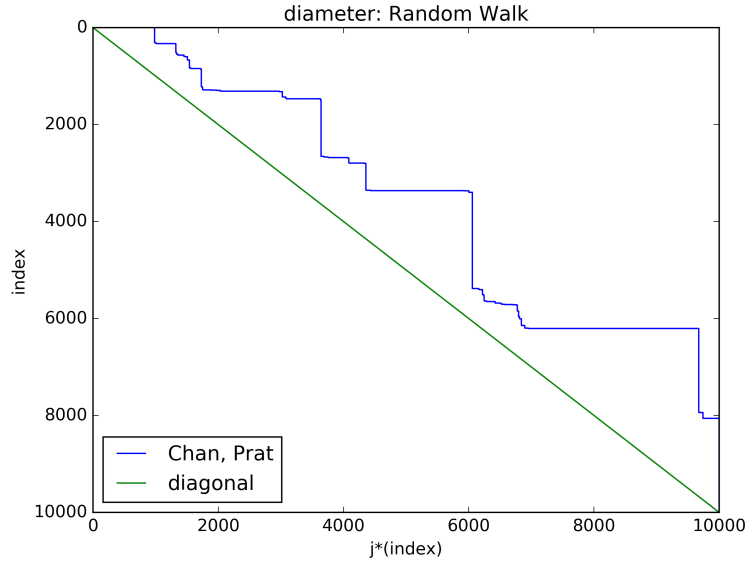


Figure 4: Caption...

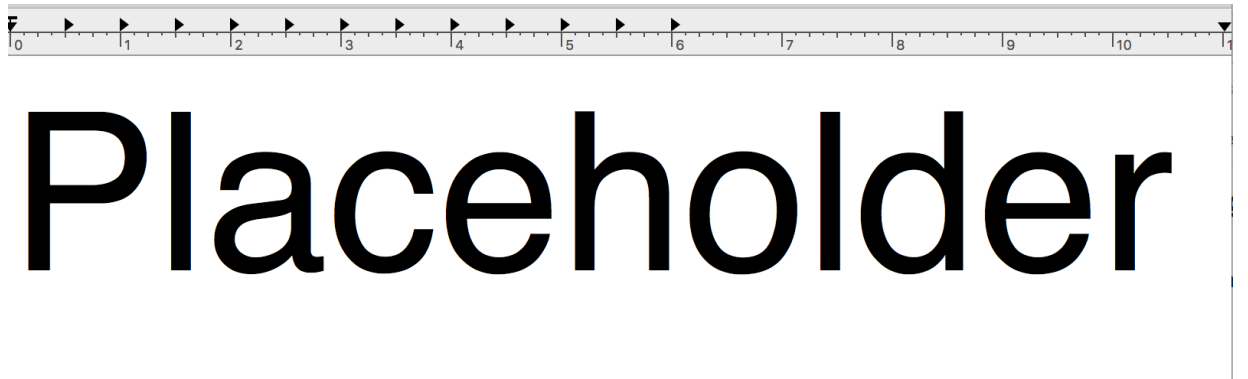


Figure 5: Runtimes (in milliseconds) of Naive and Chan, Prat [Similar to Figure 3]

References

- [1] Drago Bokal, Sergio Cabello, and David Eppstein. Finding All Maximal Subsequences with Hereditary Properties. In Lars Arge and János Pach, editors, *31st International Symposium on Computational Geometry (SoCG 2015)*, volume 34 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 240–254, Dagstuhl, Germany, 2015. Schloss Dagstuhl–

Leibniz-Zentrum fuer Informatik.

- [2] Timothy M. Chan and Simon Pratt. Two Approaches to Building Time-Windowed Geometric Data Structures. In Sándor Fekete and Anna Lubiw, editors, *32nd International Symposium on Computational Geometry (SoCG 2016)*, volume 51 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:15, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.