Clustering
with the
Dynamic Time Warping Distance

Koen van Greevenbroek

Master’s Thesis Seminar, 30 March 2020

Outline

Review

The DTW distance and k-Median Problem.

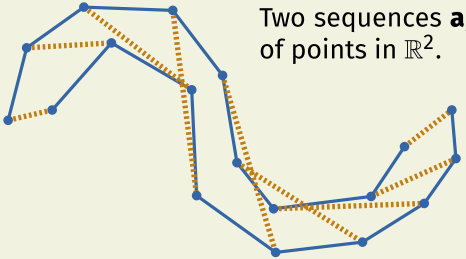
Understanding solutions

Exact algorithms and heuristics.

Better approximation algorithms

Approaches and obstacles.

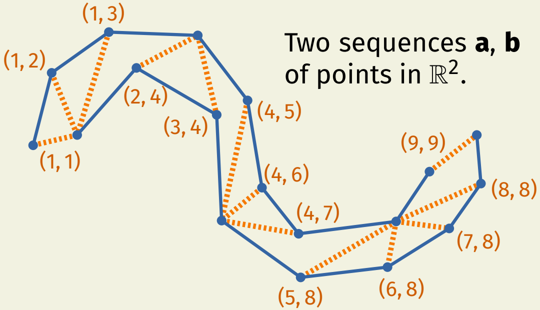
Distance between Sequences



Two sequences **a**, **b** of points in $\mathbb{R}^2$.

Naïve distance: $\sum_i d(a_i, b_i)$ or $\sqrt{\sum_i d(a_i, b_i)^2}$.

Distance between Sequences



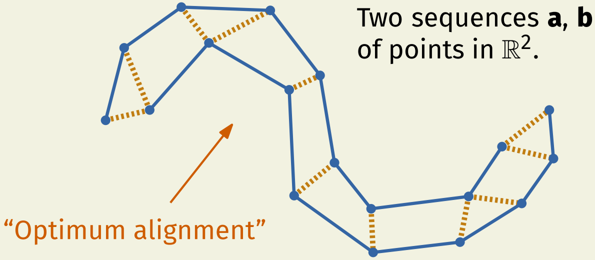
Two sequences **a**, **b** of points in $\mathbb{R}^2$.

Alignment between two sequences:

$$(1, 1) = (i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N) = (m, m)$$

such that the *i*s and *j*s are non-decreasing, and $i_{k+1} - i_k \leq 1$ and $j_{k+1} - j_k \leq 1$ for all k.

Distance between Sequences

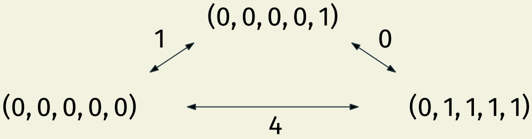


Dynamic Time Warping (DTW) distance: "Minimum cost of matching up the points of the sequences."

$$\mathrm{DTW}(\mathbf{a}, \mathbf{b}) = \min_{\mathcal{A} \text{ an alignment}} \sum_{(i,j)\in\mathcal{A}} d(a_i, b_j).$$

DTW: Recap

- We can compute DTW($\mathbf{a}, \mathbf{b}$) in $O(n^2)$ with a dynamic program.
- The DTW distance is similar to the discrete Fréchet distance ($\sum$ instead of max).
- The DTW distance does *not* satisfy the triangle-inequality:

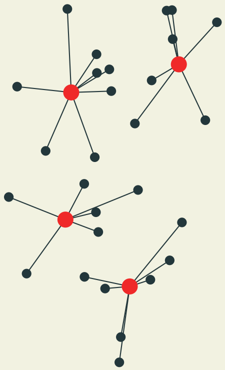


The k-Median Problem

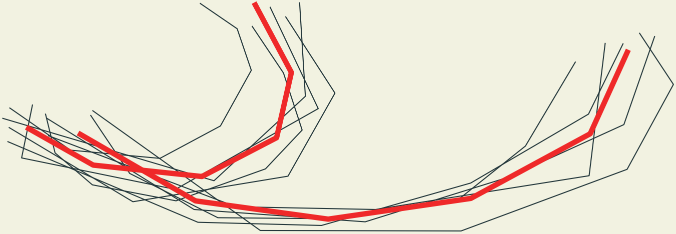
Let (X, D) be a set with a distance function. For a subset $P \subseteq X$, find k points $C \subseteq X$ which minimize $\sum_{p \in P} \min_{c \in C} D(c, p)$.

Facts:

- Most variations of the k-Median Problem are NP-hard.
- *Over discrete metric spaces:* constant factor approximation, but no $(1 + \epsilon)$-approximation algorithm.
- *Over* $\mathbb{R}^d$*:* cannot compute exactly, but there are good $(1 + \epsilon)$-approximation algorithms.



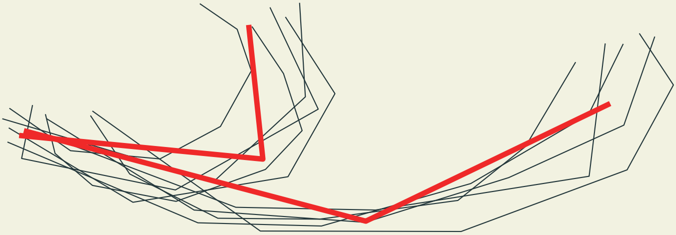
Goal: Study the DTW k-Median Problem.



Fact: The DTW 1-Median Problem is NP-hard.

- Exponential time exact algorithms?
- Constant factor approximation algorithms?
- Hard to approximate?

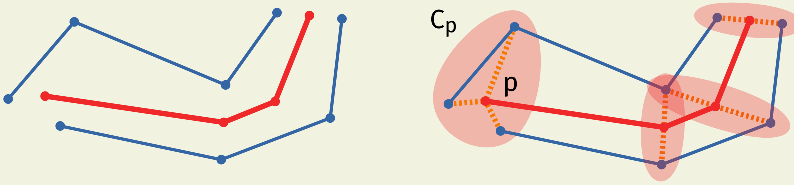
Goal: Study the DTW (k, ℓ)-Median Problem.



Fact: The DTW $(1, \ell)$-Median Problem is NP-hard.

- Exponential time exact algorithms?
- Constant factor approximation algorithms?
- Hard to approximate?

DTW 1-Median Problem: structure of optimum centre curves

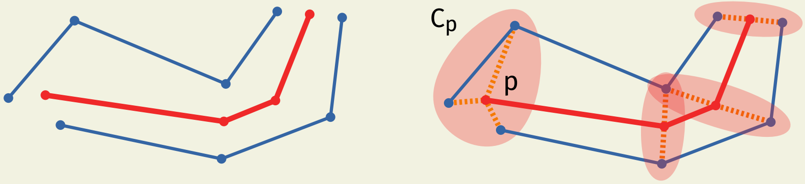


Let C_p be the points on the input curves matched to a point p on the centre curve by optimum alignments.

Proposition: For p a point on an optimum centre curve, p is the 1-median of C_p in the underlying space X.

Proof: If not, we could replace p by the 1-median of C_p in X to get a better centre curve.

DTW 1-Median Problem: structure of optimum centre curves



Consequence:

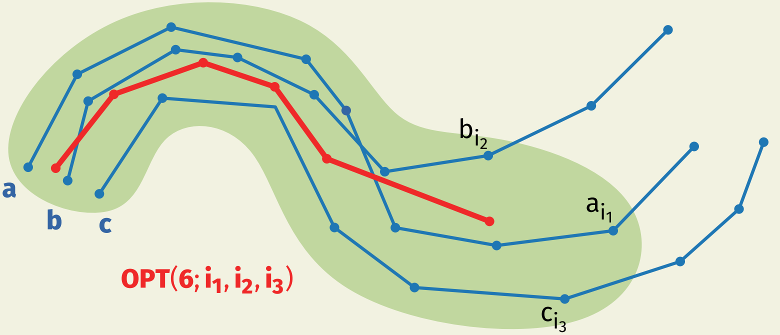
Even if the underlying space X is infinite (e.g. $X = \mathbb{R}^d$), there is a finite set of potential centre curves: those using points that are 1-medians of subsets of the points of the input curves.

Note: need to compute or approximate 1*-medians in X.*

An exact dynamic program

n: number of curves

m: length of curve



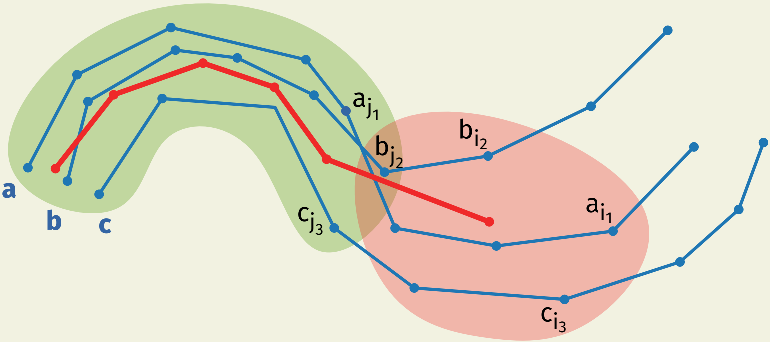
Let $\mathrm{OPT}(\ell; i_1, i_2, \ldots, i_n)$ be an optimum centre curve of length ℓ for the input curves truncated to the $i_1, i_2, \ldots, i_n$th points, respectively.

The dynamic programming table is of size m^{n+1}.

An exact dynamic program

n: number of curves

m: length of curve



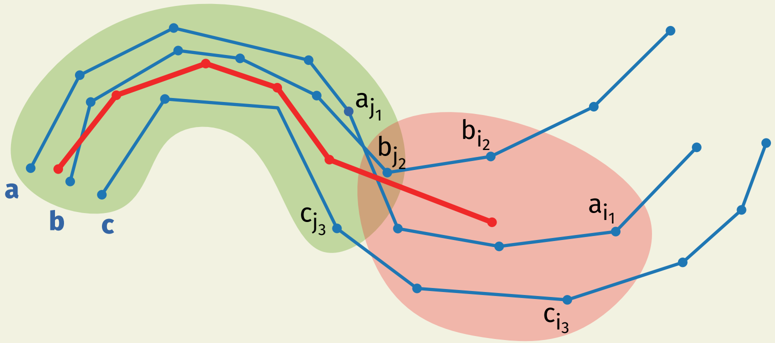
To compute $\mathrm{OPT}(\ell; i_1, \ldots, i_n)$, use the previous solutions $\mathrm{OPT}(\ell - 1; j_1, \ldots, j_n)$ for all $j_1 \leq i_1, \ldots, j_n \leq i_n$.

The last point on $\mathrm{OPT}(\ell, i_1, i_2, i_3)$ is the 1-median of $a_{j_1+1}, \ldots, a_{i_1}, b_{j_2+1}, \ldots, b_{i_2}, c_{j_3+1}, \ldots, c_{i_3}$, and possibly a_{j_1}, b_{j_2} and c_{j_3}. *(Case of $n = 3$ as above.)*

An exact dynamic program

n: number of curves

m: length of curve



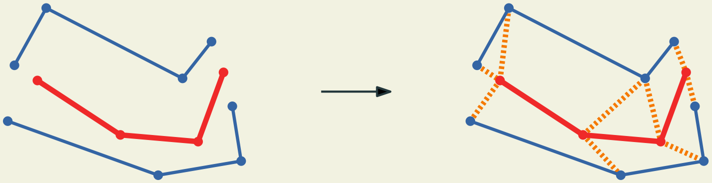
The resulting dynamic programming algorithm has a time complexity of $O(m^{2n+3}2^n n)$.

Is there an exact algorithm with runtime close to $O(m^n)$?

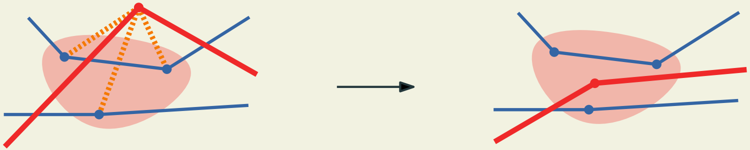
A heuristic

There is a useful local optimization heuristic similar to the k-means algorithm, called the *DTW Barycentric Average (DBA)* algorithm. Repeat the following two steps:

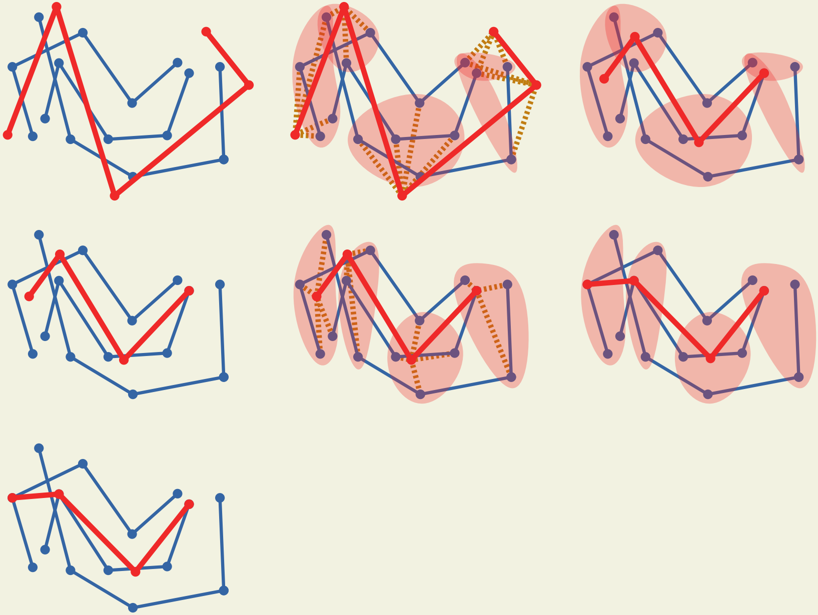
1. **Align:** find optimum alignments.



2. **Refine:** set each p to the 1-median of C_p.



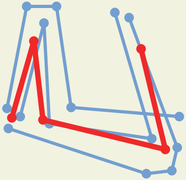
Example:



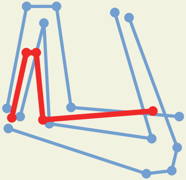
The DBA heuristic: facts

- Widely used in practice: good results, fast, easy.
- A family of examples show that the DBA algorithm does *not* have a constant factor approximation ratio.

Optimum centre curve:



DBA gets stuck here:



Are there “good” ways to initialize the DBA heuristic?

Better approximation algorithms

Difficulty in finding an optimum centre curve:

Points: There are up to $(m(m+1)/2)^n$ potential centre curve points: 1-medians of subsets of points on the input curves that could be used for a centre curve.

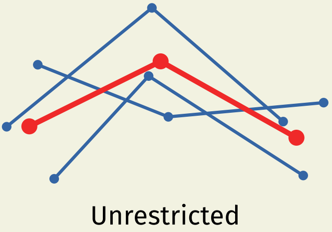
Order: There are N^m ordered sequences of length m on any N distinct points.

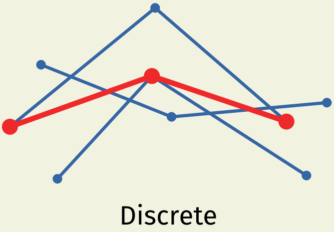
Note:

- The number of points can be limited.
- The NP-hardness proof for the DTW 1-Median Problem uses only the points $\{-1, 0, 1\}$: the *order* is the hardest part.

Discrete solutions

A *discrete* centre curve is restrict to only using points from the input curves.



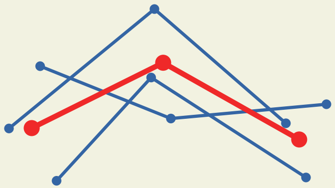


This is a 2-approx. when the underlying space is metric.

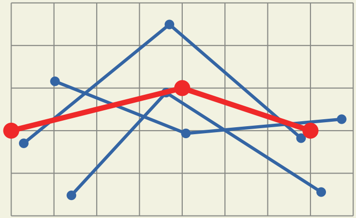
Reduces the number of potential centre curve points from $(m(m+1)/2)^n$ to mn.

Solutions on a grid (for $X = \mathbb{R}^d$)

A *grid* centre curve is restrict to only using points on a grid with a given resolution.



Unrestricted

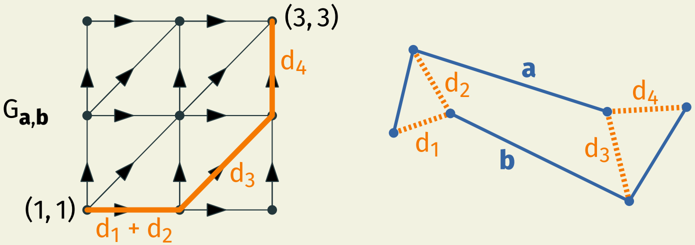


Restricted to grid

Used to develop $(1+\epsilon)$-approximation algorithms for the $(1,\ell)$-Median Problem w.r.t. the discrete Fréchet distance, for constant ℓ. May make the number of potential centre curve points *independent* of *n*.

Linear programming formulation

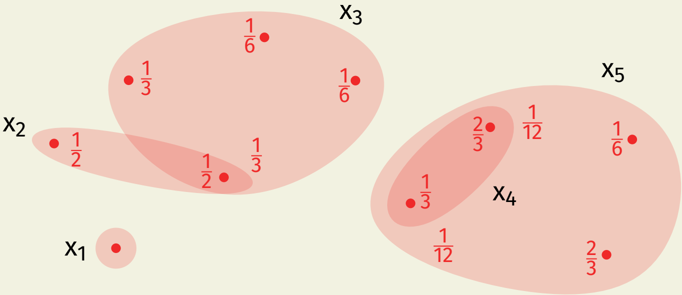
We can write the DTW distance between curves **a** and **b** as a linear program. Let $G_{\mathbf{a},\mathbf{b}}$ be an $m \times m$ grid graph with all directed edges makes steps $(0,1)$, $(1,0)$ and $(1,1)$. Cost of edges into (i,j) is $d(a_i, b_j)$. Add cost of $d(a_1, b_1)$ to outgoing edges from $(1,1)$. Let $(1,1)$ be a source and (m,m) a sink of value 1.



$$\mathrm{DTW}(\mathbf{a}, \mathbf{b}) = \min\{c(f) \mid f \text{ a flow in } G_{\mathbf{a},\mathbf{b}}\}.$$

Linear programming formulation

A fractional curve **x** on a set of points P given by variables x_{ia} for $1 \leq i \leq m$ and $1 \leq a \leq |P|$ with $\sum_{a=1}^{|P|} x_{ia} = 1$ for all i:



We can define the DTW distance between **x** and a normal curve **b**. Define a graph $G_{\mathbf{x},\mathbf{b}}$ analogous to $G_{\mathbf{a},\mathbf{b}}$ but on a $|P| \times m \times m$ grid, and let

$$\mathrm{DTW}(\mathbf{x}, \mathbf{b}) = \min\{c(f) \mid f \text{ a flow in } G_{\mathbf{x},\mathbf{b}}\}.$$

Linear programming formulation

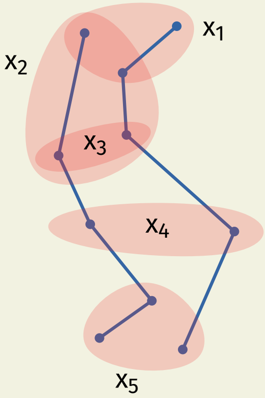
Let $\mathcal{S}$ be n curves. A relaxation of the Discrete 1-Median Problem on $\mathcal{S}$ is

$$\min \left\{ \sum_{\mathbf{s} \in \mathcal{S}} c(f_{\mathbf{s}}) \;\middle|\; f_{\mathbf{s}} \text{ a flow in } G_{\mathbf{x},\mathbf{s}},\ 0 \le x_{ia} \le 1 \right\}.$$

Although we can solve the above LP in polynomial time, the integrality ratio is unbounded.

Example

Input sequences: $(0, 1, 0, 1, 0, 0)$ and $(0, 1, 0, 0, 0, 0)$. The fractional centre curve $(0, \{0, 1\}, \{0, 1\}, \{0, 1\}, \{0, 1\}, 0)$ has cost 0, but there is no integral centre curve of cost 0.



Questions?