Coresets for (k, ℓ)-Median under Dynamic Time Warping

Jacobus Conradi, Benedikt Kolbe, Ioannis Psarros, Dennis Rohde

UNIVERSITÄT BONN

conradi@cs.uni-bonn.de

EuroCG'24, Ioannina

Prelude

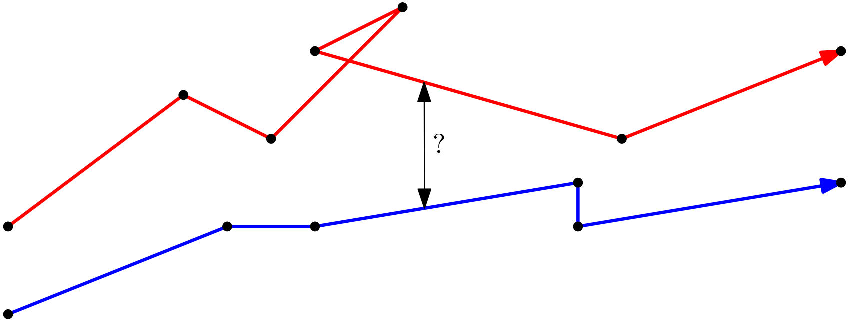
The key players

Key player I – curves of complexity $\leq m$

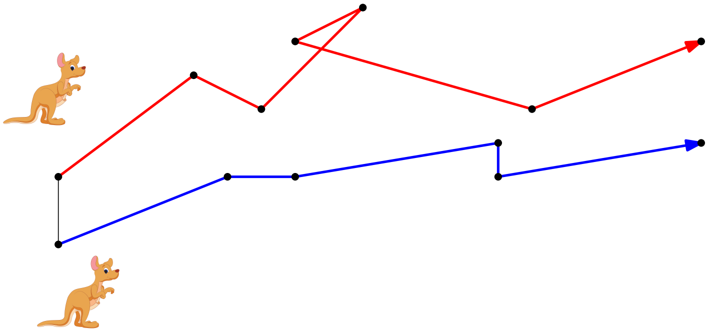
Central objects $\mathbb{X}^d_m$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}^d_m$: point sequences $\tau = (\tau_1, \ldots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.

Distance measure: dynamic time warping (dtw)

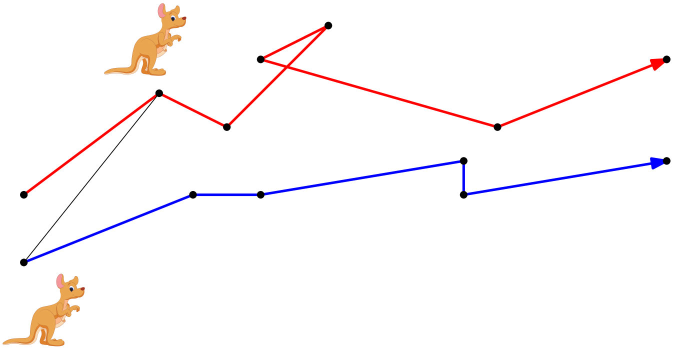


Dynamic Time Warping



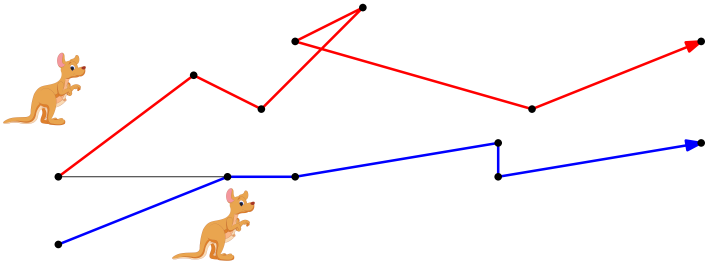
joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



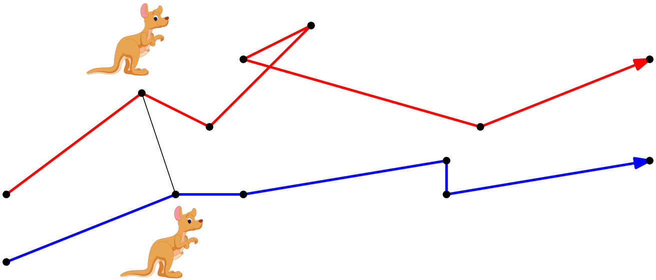
joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



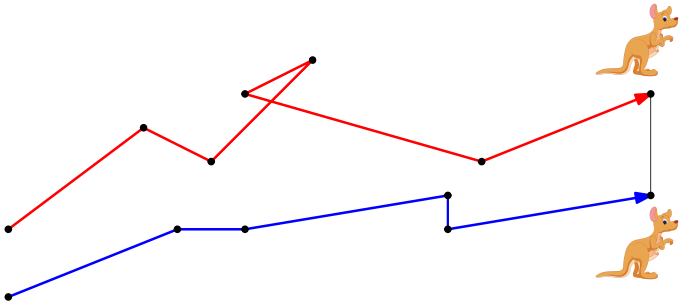
joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



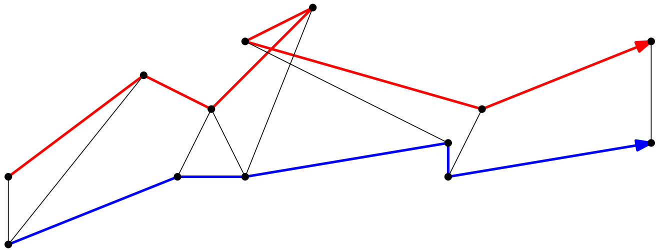
joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



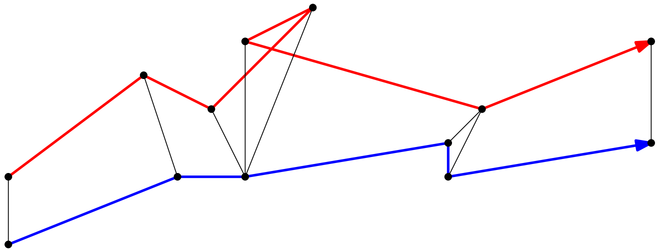
joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.

Dynamic Time Warping



joint traversals: Each kangaroo allowed to hop either one step ahead, or stay put.
dtw: minimal possible sum of Euclidean distances.

Key player I – curves of complexity $\leq m$ under dtw

Central objects $\mathbb{X}_m^d$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}_m^d$: point sequences $\tau = (\tau_1, \ldots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.
Distance measure: dynamic time warping (dtw)

Definition (DTW)

Let $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathbb{X}^d_{=m}, \tau = (\tau_1, \ldots, \tau_\ell) \in \mathbb{X}^d_{=\ell}$. The DTW of σ and τ is

$$\mathrm{dtw}(\sigma, \tau) = \min_{T \in \mathcal{T}_{m,\ell}} \sum_{(i,j) \in T} \|\sigma_i - \tau_j\|_2 \quad .$$

- dtw less sensitive to outliers than Fréchet but not a metric
- $\mathrm{dtw}_p \xrightarrow{p \to \infty} d_{\text{discrete Fréchet}}$; $\mathrm{dtw}_1 = \mathrm{dtw}$

Key player I – curves of complexity $\leq m$ under dtw

Central objects $\mathbb{X}^d_m$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}^d_m$: point sequences $\tau = (\tau_1, \ldots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.
Distance measure: dynamic time warping (dtw)

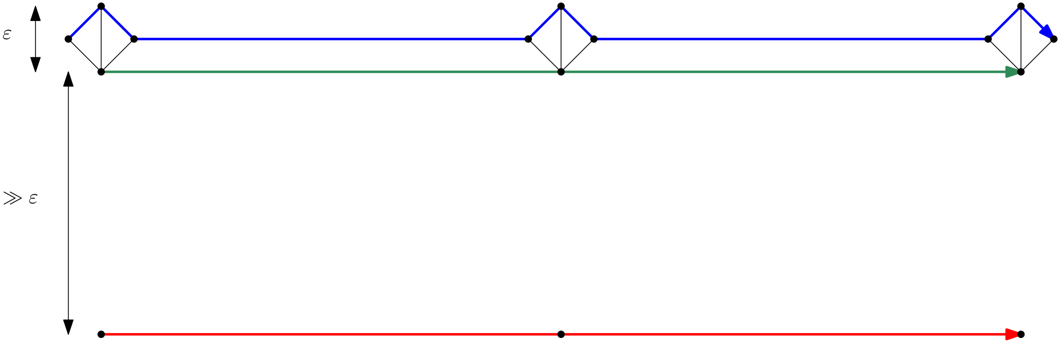
Definition (DTW)

Let $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathbb{X}^d_{=m}, \tau = (\tau_1, \ldots, \tau_\ell) \in \mathbb{X}^d_{=\ell}$. The DTW of σ and τ is

$$\mathrm{dtw}(\sigma, \tau) = \min_{T \in \mathcal{T}_{m,\ell}} \sum_{(i,j) \in T} \|\sigma_i - \tau_j\|_2 \quad .$$

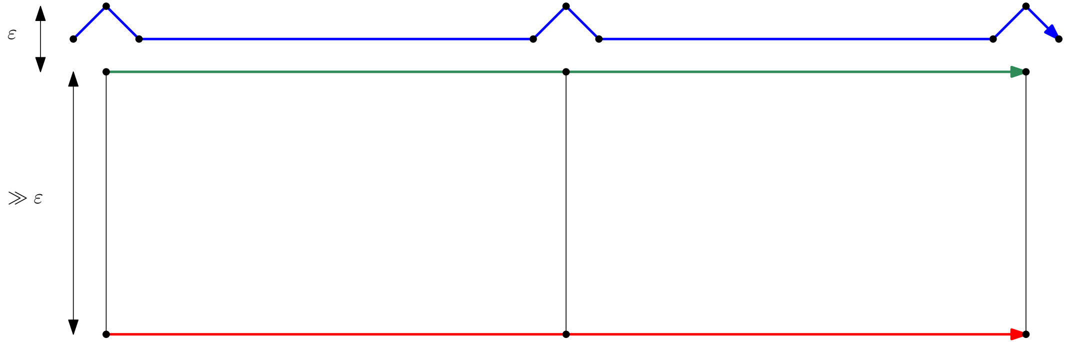
- dtw less sensitive to outliers than Fréchet but not a metric
- $\mathrm{dtw}_p \xrightarrow{p \to \infty} d_{\text{discrete Fréchet}}$; $\mathrm{dtw}_1 = \mathrm{dtw}$

Vioalted triangle inequality



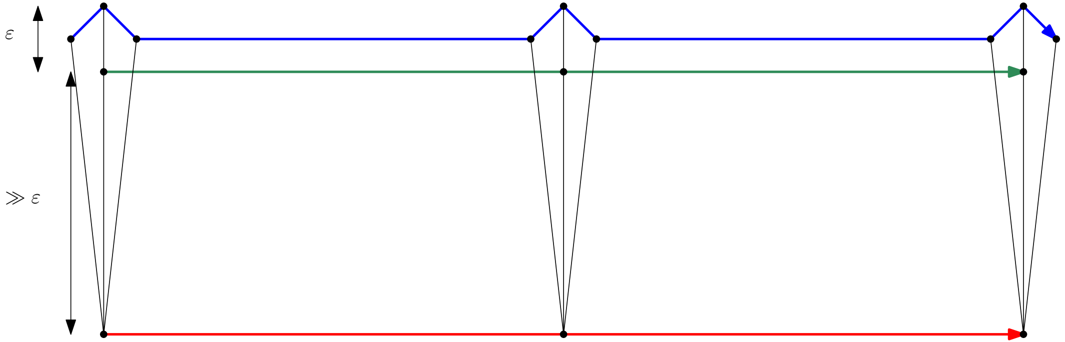
dtw(■, ■) $\approx 9\epsilon$

Vioalted triangle inequality



$\text{dtw}(\blacksquare, \blacksquare) \approx 3$

Vioalted triangle inequality



dtw(■, ■) ≈ 9 but $9 \not\leq 3 + 9\epsilon$

Key player I – curves of complexity $\leq m$ under dtw

Central objects $\mathbb{X}^d_m$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}^d_m$: point sequences $\tau = (\tau_1, \ldots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.
Distance measure: dynamic time warping (dtw)

Definition (DTW)

Let $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathbb{X}^d_{=m}, \tau = (\tau_1, \ldots, \tau_\ell) \in \mathbb{X}^d_{=\ell}$. The DTW of σ and τ is

$$\mathrm{dtw}(\sigma, \tau) = \min_{T \in \mathcal{T}_{m,\ell}} \sum_{(i,j) \in T} \|\sigma_i - \tau_j\|_2 \quad .$$

- dtw less sensitive to outliers than Fréchet but not a metric
- $\mathrm{dtw}_p \xrightarrow{p \to \infty} d_{\text{discrete Fréchet}}$; $\mathrm{dtw}_1 = \mathrm{dtw}$

Key player I – curves of complexity $\leq m$ under dtw

Central objects $\mathbb{X}^d_m$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}^d_m$: point sequences $\tau = (\tau_1, \ldots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.
Distance measure: dynamic time warping (dtw)

Definition (p-DTW)

Let $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathbb{X}^d_{=m}, \tau = (\tau_1, \ldots, \tau_\ell) \in \mathbb{X}^d_{=\ell}$. The p-DTW of σ and τ is

$$\mathrm{dtw_p}(\sigma,\tau) = \min_{T \in \mathcal{T}_{m,\ell}} \left(\sum_{(i,j)\in T} \|\sigma_i - \tau_j\|_2^p \right)^{1/p}$$ for $p \in [1, \infty)$.

- dtw less sensitive to outliers than Fréchet but not a metric
- $\mathrm{dtw_p} \xrightarrow{p\to\infty} \mathrm{d_{discrete\ Fréchet}}$; $\mathrm{dtw_1} = \mathrm{dtw}$

Key player I – curves of complexity $\leq m$ under dtw

Central objects $\mathbb{X}^d_m$ and dtw

Complexity $\leq m$ (polygonal) **curves** in $\mathbb{X}^d_m$: point sequences $\tau = (\tau_1, \dots, \tau_{m'})$ with $\tau_i \in \mathbb{R}^d$ and $m' \leq m$.

Distance measure: dynamic time warping (dtw)

Definition (p-DTW)

Let $\sigma = (\sigma_1, \dots, \sigma_m) \in \mathbb{X}^d_{=m}, \tau = (\tau_1, \dots, \tau_\ell) \in \mathbb{X}^d_{=\ell}$. The p-DTW of σ and τ is $\mathrm{dtw_p}(\sigma, \tau) = \min_{T \in \mathcal{T}_{m,\ell}} \left(\sum_{(i,j) \in T} \|\sigma_i - \tau_j\|_2^p \right)^{1/p}$ for $p \in [1, \infty)$.

- dtw less sensitive to outliers than Fréchet but not a metric
- $\mathrm{dtw_p} \xrightarrow{p \to \infty} \mathrm{d_{discrete\ Fréchet}}$; $\mathrm{dtw_1} = \mathrm{dtw}$

k-median problem for curves

(k, ℓ)-median problem for $\mathbb{X}^d_m$ and $k, \ell \in \mathbb{N}$:

Given a set of $n \in \mathbb{N}$ curves $T = \{\tau_1, \ldots, \tau_n\} \subset \mathbb{X}^d_m$, identify k center curves $C = \{c_1, \ldots, c_k\} \subset \mathbb{X}^d_\ell$ that minimize $\text{cost}(T, C) = \sum_{\tau \in T} \min_{c \in C} \text{dtw}_p(\tau, c)$.

k-median problem for curves

(k, ℓ)-median problem for $\mathbb{X}^d_m$ and $k, \ell \in \mathbb{N}$:

Given a set of $n \in \mathbb{N}$ curves $T = \{\tau_1, \dots, \tau_n\} \subset \mathbb{X}^d_m$, identify k center curves $C = \{c_1, \dots, c_k\} \subset \mathbb{X}^d_\ell$ that minimize $\mathrm{cost}(T, C) = \sum_{\tau \in T} \min_{c \in C} \mathrm{dtw_p}(\tau, c)$.

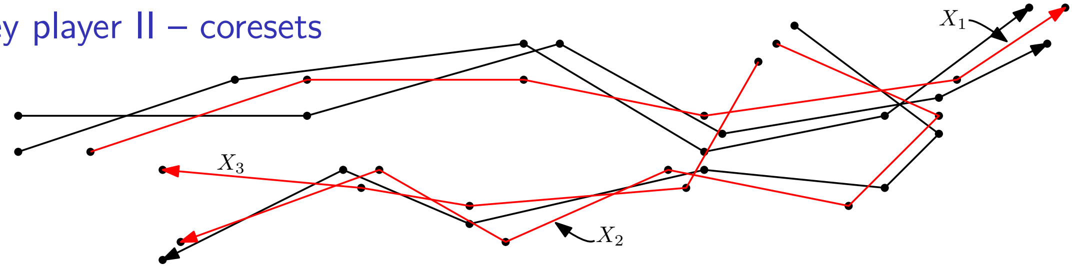
Definition ((α, β)-approximation for (k, ℓ)-median)

A set $\hat{C} \subset \mathbb{X}^d_\ell$ is an (α, β)-approximation of (k, ℓ)-median if $|\hat{C}| \leq \boldsymbol{\beta} k$ and

$$\mathrm{cost}\left(T, \hat{C}\right) \leq \boldsymbol{\alpha}\, \mathrm{cost}(T, C) = \boldsymbol{\alpha} \sum_{\tau \in T} \min_{c \in C} \mathrm{dtw_p}(\tau, c)$$

for any $C \subset \mathbb{X}^d_\ell$ of size k.

Key player II – coresets



Definition (ϵ-coreset)

Let $T \subset \mathbb{X}^d_m$ and $\epsilon \in (0, 1)$. A weighted multiset $S \subset \mathbb{X}^d_m$ with weights $w : S \to \mathbb{R}_{>0}$ is a ϵ-**coreset** for (k, ℓ)-median of T under $\mathrm{dtw_p}$ if $\forall C \subset \mathbb{X}^d_\ell$ with $|C| = k$

$$(1 - \epsilon)\,\mathrm{cost}(T, C) \leq \sum_{s \in S} w(s) \min_{c \in C} \mathrm{dtw_p}(s, c) \leq (1 + \epsilon)\,\mathrm{cost}(T, C).$$

Part I

Sensitivity-sampling

or

Identifying our problems

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho \log(n))$

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho \log(n))$

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho \log(n))$

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k \rho \log(n))$

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho\log(n))$

Problems with $\mathrm{dtw_p}$:

- No known polynomial (α, β)-approximation algorithm for (k, ℓ)-median
- VC dimension of $\mathrm{dtw_p}$ is not known
- $\mathrm{dtw_p}$ is no metric

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho \log(n))$

Problems with dtw_p:

- No known polynomial (α, β)-approximation algorithm for (k, ℓ)-median
- VC dimension of dtw_p is not known
- dtw_p is no metric

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho\log(n))$

Problems with dtw_p**:**

- No known polynomial (α, β)-approximation algorithm for (k, ℓ)-median
- VC dimension of dtw_p is not known
- dtw_p is no metric

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α,β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho\log(n))$

Problems with $\mathrm{dtw_p}$:

- No known polynomial (α,β)-approximation algorithm for (k,ℓ)-median
- VC dimension of $\mathrm{dtw_p}$ is not known
- $\mathrm{dtw_p}$ is no metric

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho\log(n))$

Problems with $\mathrm{dtw_p}$:

- No known polynomial (α, β)-approximation algorithm for (k, ℓ)-median
- ~~VC dimension of $\mathrm{dtw_p}$ is not known~~
- $\mathrm{dtw_p}$ is no metric

Theorem (C, Kolbe, Psarros, Rohde)

There is a distance $\widetilde{\mathrm{dtw_p}}$ with $\mathrm{dtw_p} \leq \widetilde{\mathrm{dtw_p}} \leq (1+\varepsilon)\,\mathrm{dtw_p}$. Its associated ball range space has VC dimension $O(d\ell\log(\ell m\varepsilon^{-1}))$.

- $\widetilde{\mathrm{dtw_p}}$ is hard-to-compute so computation is circumvented

Sensitivity ($\widehat{=}$importance) sampling

Framework for ϵ-coreset for k-median in metric spaces [Feldman,Langberg'11,Braverman,Feldman,Lang'16,...]

- If there is a (α, β)-approximation algorithm and
- If the metric space has VC dimension ρ
- Then we can sample a coreset of size roughly $\widetilde{O}(\epsilon^{-2}\alpha\beta k\rho\log(n))$

Problems with $\mathrm{dtw_p}$:

- No known polynomial (α, β)-approximation algorithm for (k, ℓ)-median
- ~~VC dimension of $\mathrm{dtw_p}$ is not known~~
- $\mathrm{dtw_p}$ is no metric

Theorem (C, Kolbe, Psarros, Rohde)

There is a distance $\widetilde{\mathrm{dtw_p}}$ *with* $\mathrm{dtw_p} \leq \widetilde{\mathrm{dtw_p}} \leq (1+\varepsilon)\,\mathrm{dtw_p}$. *Its associated ball range space has VC dimension* $O(d\ell\log(\ell m\varepsilon^{-1}))$.

- $\widetilde{\mathrm{dtw_p}}$ is hard-to-compute so computation is circumvented

Part II

(α, β)-approximations of (k, ℓ)-median

or

How to force DTW to be a metric

The iterated triangle inequality

Lemma (Iterated triangle inequality)

Let $s \in \mathbb{X}^d_\ell$, $t \in \mathbb{X}^d_{\ell'}$ and $X = (x_1, \ldots, x_r)$ be any ordered set of curves in $\mathbb{X}^d_m$. Then

$$\mathrm{dtw_p}(s,t) \leq (\ell + \ell')^{1/p} \left(\mathrm{dtw_p}(s, x_1) + \sum_{i<r} \mathrm{dtw_p}(x_i, x_{i+1}) + \mathrm{dtw_p}(x_r, t) \right).$$

- No dependence on complexities of visited curves
- Connection to empirical observation that the Δ-inequality is only rarely violated

The iterated triangle inequality

Lemma (Iterated triangle inequality)

Let $s \in \mathbb{X}^d_\ell$, $t \in \mathbb{X}^d_{\ell'}$ and $X = (x_1, \ldots, x_r)$ be any ordered set of curves in $\mathbb{X}^d_m$. Then

$$\mathrm{dtw_p}(s,t) \leq (\ell + \ell')^{1/p} \left(\mathrm{dtw_p}(s, x_1) + \sum_{i<r} \mathrm{dtw_p}(x_i, x_{i+1}) + \mathrm{dtw_p}(x_r, t) \right).$$

- No dependence on complexities of visited curves
- Connection to empirical observation that the Δ-inequality is only rarely violated

The iterated triangle inequality

Lemma (Iterated triangle inequality)

Let $s \in \mathbb{X}^d_\ell$, $t \in \mathbb{X}^d_{\ell'}$ and $X = (x_1, \dots, x_r)$ be any ordered set of curves in $\mathbb{X}^d_m$. Then

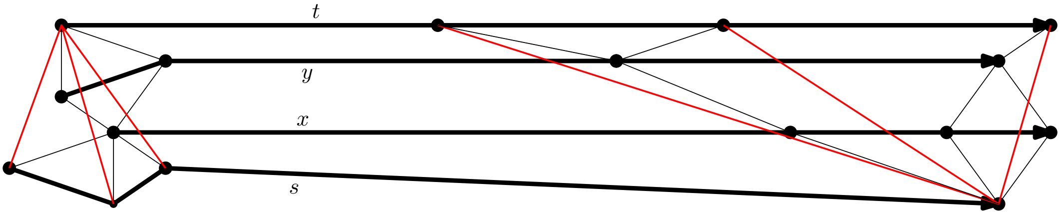
$$\mathrm{dtw_p}(s,t) \leq (\ell + \ell')^{1/p} \left(\mathrm{dtw_p}(s, x_1) + \sum_{i<r} \mathrm{dtw_p}(x_i, x_{i+1}) + \mathrm{dtw_p}(x_r, t) \right).$$

- No dependence on complexities of visited curves
- Connection to empirical observation that the Δ-inequality is only rarely violated

The iterated triangle inequality illustrated

Key observation

Proof of Δ-inequality boils down to constructing some traversal between the end curves *s* and *t* from optimal traversals for the inbetween curves.



Bridging the gap to a metric space

Consequences of iterated Δ-inequality

- Shortest path from σ to τ in T not necessarily direct edge between them
- However, the direct edge is not too far off
- Defining $\overline{\mathrm{dtw_p}|_T}$ to be the **metric closure** (shortest path metric), we have

$$\overline{\mathrm{dtw_p}|_T} \leq \mathrm{dtw_p}|_T \leq (2m)^{1/p}\overline{\mathrm{dtw_p}|_T} \leq (2m)^{1/p}\,\mathrm{dtw_p}|_T$$

Bridging the gap to a metric space

Consequences of iterated Δ-inequality

- Shortest path from σ to τ in T not necessarily direct edge between them
- However, the direct edge is not too far off
- Defining $\overline{\mathrm{dtw_p}|_T}$ to be the **metric closure** (shortest path metric), we have

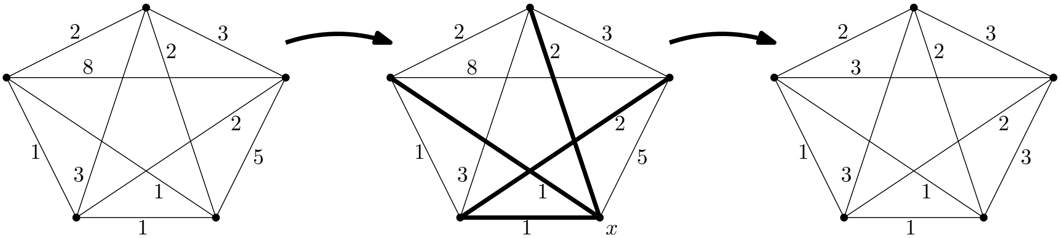
$$\overline{\mathrm{dtw_p}|_T} \leq \mathrm{dtw_p}|_T \leq (2m)^{1/p}\overline{\mathrm{dtw_p}|_T} \leq (2m)^{1/p}\,\mathrm{dtw_p}|_T$$

Bridging the gap to a metric space

Consequences of iterated Δ-inequality

- Shortest path from σ to τ in T not necessarily direct edge between them
- However, the direct edge is not too far off
- Defining $\overline{\mathrm{dtw_p}\,|_T}$ to be the **metric closure** (shortest path metric), we have

$$\overline{\mathrm{dtw_p}\,|_T} \leq \mathrm{dtw_p}\,|_T \leq (2m)^{1/p}\overline{\mathrm{dtw_p}\,|_T} \leq (2m)^{1/p}\,\mathrm{dtw_p}\,|_T$$



Denouement - putting it all together

We can now access the wealth of k-median results in metric spaces:

Theorem (C, Kolbe, Psarros, Rohde)

We can compute a $(O((1+\epsilon)(m\ell)^{1/p}), 4)$*-approximation for* (k, ℓ)*-median in* $\mathbb{X}_m^d$ *in time*

$$\widetilde{O}(n \cdot \mathrm{poly}(m, \ell, k, d, \epsilon^{-1}))$$

Denouement - putting it all together

We can now access the wealth of k-median results in metric spaces:

Theorem (C, Kolbe, Psarros, Rohde)

We can compute a $(O((1+\epsilon)(m\ell)^{1/p}), 4)$-approximation for (k, ℓ)-median in $\mathbb{X}_m^d$ in time

$$\widetilde{O}(n \cdot \mathrm{poly}(m, \ell, k, d, \epsilon^{-1}))$$

With the last dtw_p-problem adressed:

Denouement - putting it all together

We can now access the wealth of k-median results in metric spaces:

Theorem (C, Kolbe, Psarros, Rohde)

We can compute a $(O((1+\epsilon)(m\ell)^{1/p}), 4)$*-approximation for* (k, ℓ)*-median in* $\mathbb{X}_m^d$ *in time*

$$\widetilde{O}(n \cdot \mathrm{poly}(m, \ell, k, d, \epsilon^{-1}))$$

With the last $\mathrm{dtw_p}$-problem adressed:

Theorem (C, Kolbe, Psarros, Rohde)

Given a set $T \subset \mathbb{X}_m^d$ $(|T| = n)$*, sensitivity sampling yields an* ϵ*-coreset for* (k, ℓ)*-median on* T *of size (hiding other* log *factors)*

$$\widetilde{O}(\epsilon^{-2} d\ell k^2 (m\ell)^{2/p} \log(n)).$$

Take-home message

- DTW – widespread similarity measure for trajectories;
  not a metric, but not too far off
- We find new (α, β)-approximations from approximating DTW by a path metric
- We obtain bounds on the VC dimension of $\mathrm{dtw_p}$ that allow coreset constructions from approximations to (k, ℓ)-median

Take-home message

- DTW – widespread similarity measure for trajectories;
  not a metric, but not too far off
- We find new (α, β)-approximations from approximating DTW by a path metric
- We obtain bounds on the VC dimension of dtw_p that allow coreset constructions from approximations to (k, ℓ)-median

Thanks for your attention!
Questions?

Main references

- Har-Peled, S., & Sharir, M. (2011). Relative (p,ε)-Approximations in Geometry. Discrete & Computational Geometry, 45(3), 462–496.
- Buchin, M., & Rohde, D. (2022). Coresets for (k, ℓ)-Median Clustering Under the Fréchet Distance. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 13179 LNCS (pp. 167–180).
- Driemel, A., Krivosija, A., & Sohler, C. (2016). Clustering time series under the Fréchet distance. Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, 766–785.
- Buchin, M., Driemel, A., & Rohde, D. (2021). Approximating (k, ℓ)-median clustering for polygonal curves. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, 2697–2716.