Near Linear Time Approximation Schemes for Clustering of Partially Doubling Metrics

Di Yue

Joint work with

Anne Driemel
University of Bonn

Jan Höckendorff
University of Cologne

Ioannis Psarros
Athena Research Center

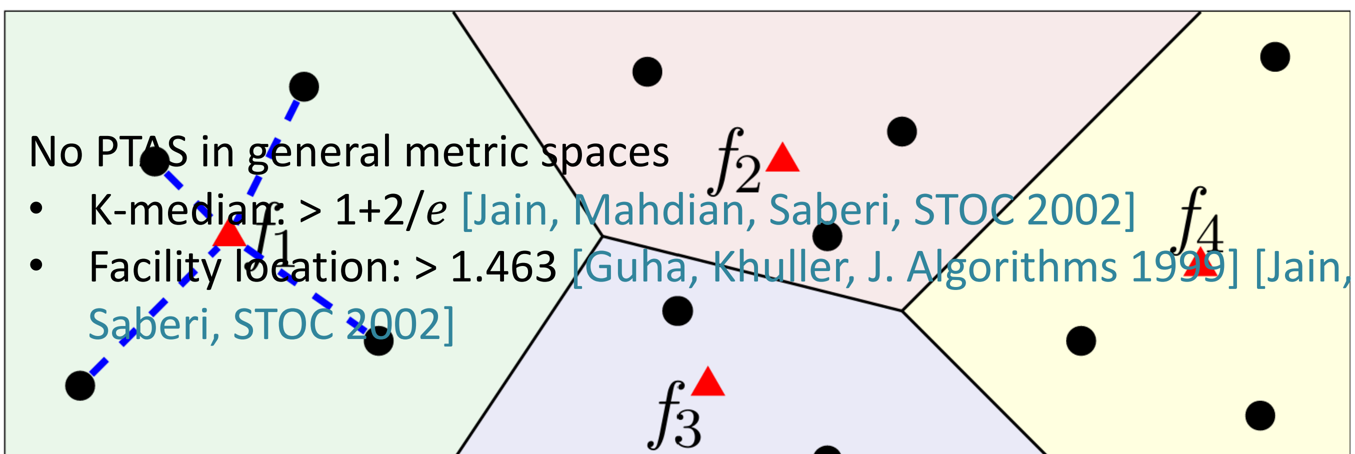
Christian Sohler
University of Cologne

Clustering

Input: finite metric space $(X \cup Y, \mathbf{d})$
- X: Clients/dataset
- Y: Candidate center/facility set

Solution: a set of centers/facilities $F \subseteq Y$

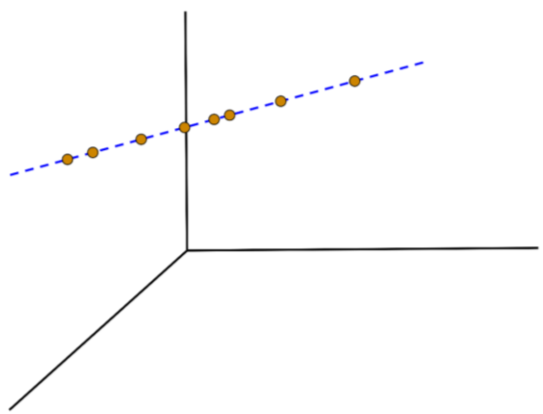
Minimize clustering objective
- K-Median: $\sum_{x\in X} \mathbf{d}(x, \mathrm{F})$ s.t. $|F| = k$
- Facility Location: $\sum_{x\in X} \mathbf{d}(x, F) + \sum_{f\in F} \mathrm{ocost}(f)$

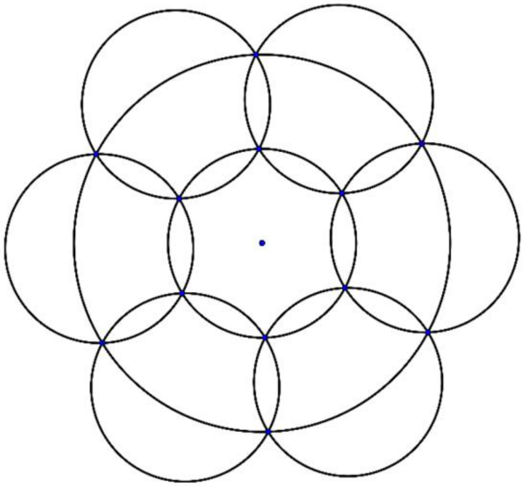


No PTAS in general metric spaces
- K-median: > 1+2/e [Jain, Mahdian, Saberi, STOC 2002]
- Facility location: > 1.463 [Guha, Khuller, J. Algorithms 1999] [Jain, Mahdian, Saberi, STOC 2002]

Doubling Dimension

Bypass the computational barrier:
Need a notion to captures the intrinsic dimension of $(X \cup Y, \mathbf{d})$





Definition (Doubling dimension of $(X \cup Y, \mathbf{d})$) [Gupta, Krauthgamer, Lee, FOCS 2003]
minimum $t > 0$, s.t. every ball in $X \cup Y$ can be covered by 2^t balls of half the radius

- $ddim(\mathbb{R}^d) = \Theta(d)$
- If $|\mathrm{P}| = n$, then $ddim(P) \le \log n$

- Natural settings of low ddim:
  - Linear subspace
  - Sparse vector representation
  - …

Clustering in Doubling Metrics

Theorem [Cohen-Addad, Feldmann, Saulpic, J. ACM 2021]
Assume $ddim(X \cup Y) \leq ddim$. Then the k-median/facility location in $(X \cup Y, \mathbf{d})$ can be $(1+\epsilon)$-approximated in time $2^{2^t}\tilde{O}(|X| + |Y|)$ with prob. 0.99, where

$$t = O\left(ddim^2 \log(1/\epsilon)\right)$$

Good news:

- Near-linear in input size
- Independent of k

Bad news:

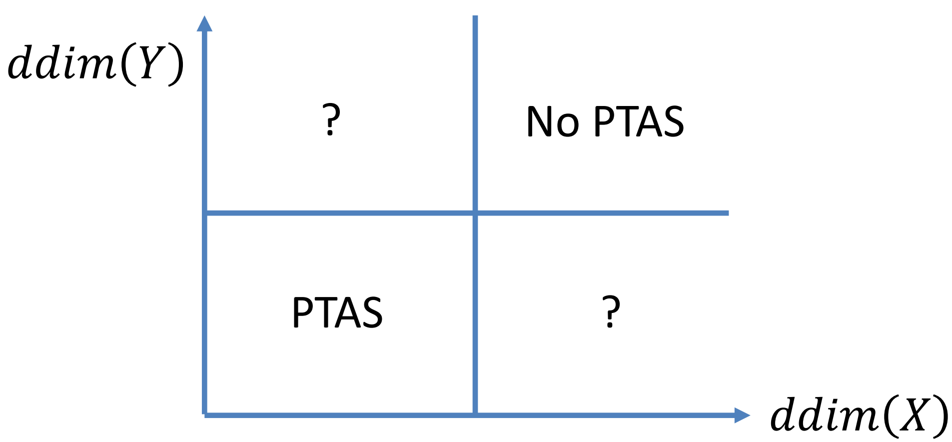
- Doubly exponential in $ddim$
    - Unavoidable by recent work [Cohen-Addad, Karthik, Saulpic, Schwiegelshohn, 2026]
- Require both X and Y have bounded ddim

Partially Doubling Metric

Theorem [Cohen-Addad, Feldmann, Saulpic, J. ACM 2021]
Assume $ddim(X \cup Y) \leq ddim$. Then the k-median/facility location in $(X \cup Y, \mathbf{d})$ can be $(1+\epsilon)$-approximated in time $2^{2^t}\tilde{O}(|X|+|Y|)$ with prob. 0.99, where

$$t = O\left(ddim^2 \log(1/\epsilon)\right)$$

Question
Does similar results hold when only one of X and Y (not both) has bounded ddim?



Partially Doubling Metric

Theorem [Cohen-Addad, Feldmann, Saulpic, J. ACM 2021]
Assume $ddim(X \cup Y) \leq ddim$. Then the k-median/facility location in $(X \cup Y, \mathbf{d})$ can be $(1+\epsilon)$-approximated in time $2^{2^t}\tilde{O}(|X| + |Y|)$ with prob. 0.99, where

$$t = O\left(ddim^2 \log(1/\epsilon)\right)$$

Question
Does similar results hold when only one of X and Y (not both) has bounded ddim?

- $ddim(X)$ bounded: Data has low-dimensional structure
- $ddim(Y)$ bounded: Centers are restricted to low dimension
  - E.g. (k, ℓ)-median problem under discrete Frechet distance [Driemel, Krivosija, Sohler, SODA 2016]

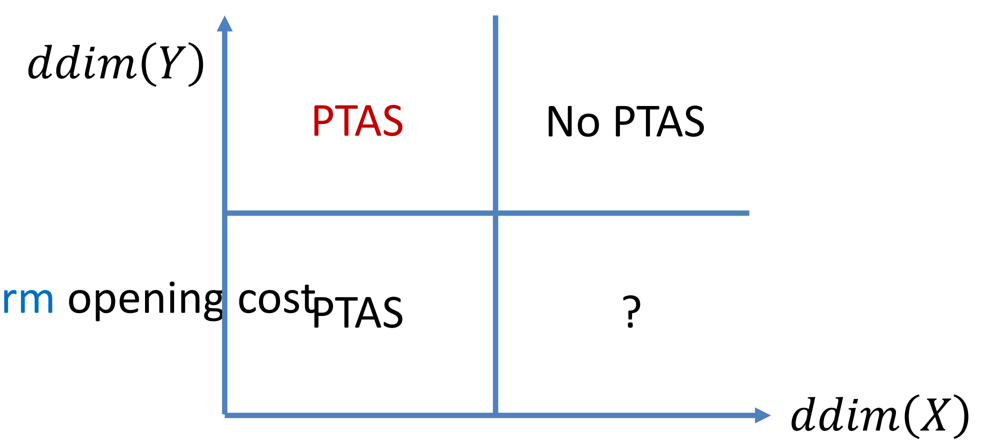
Main Results

Theorem (Low-dim clients)
Assume $ddim(\mathrm{X}) \leq ddim$. Then the k-median/facility location in $(X \cup Y, \mathbf{d})$ can be $(1+\epsilon)$-approximated in time $2^{2^t}\tilde{O}(|X|+|Y|)$ with prob. 0.99, where

$$t = O(ddim \log(\mathrm{ddim}/\epsilon))$$

[Huang, Jiang, Krauthgamer, **Y.**, STOC 2025]:
Studied a similar setting for Euclidean uniform facility location

- $Y = \mathbb{R}^d$ with $d = \Omega(\log n)$
- $X = $ doubling subset of $\mathbb{R}^d$
- Time complexity = $2^{2^t}\tilde{O}(|\mathrm{X}|)$



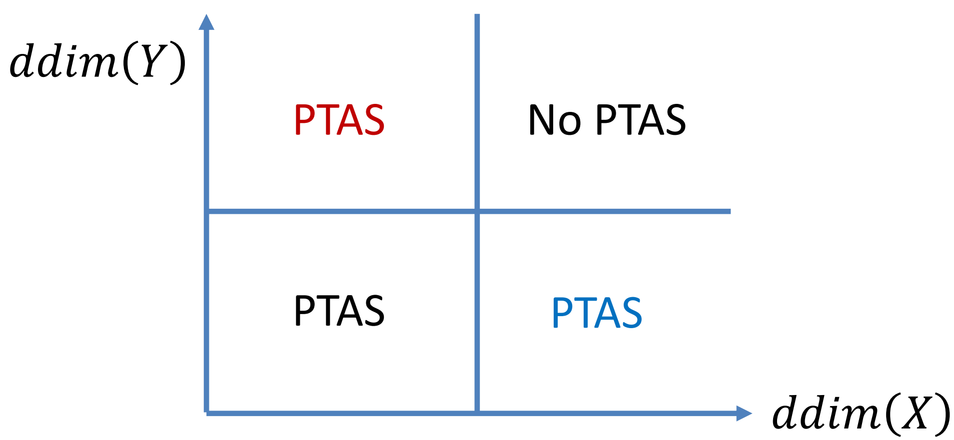
Our result is an extension

- Euclidean -> general metric spaces
- Uniform opening cost -> Non-uniform opening cost
- Also work for k-median
- But in a discrete sense: X, Y finite

Main Results

Theorem (Low-dim centers)
Assume $ddim(Y) \leq ddim$. Then the k-median/facility location in $(X \cup Y, \mathbf{d})$ can be $(1+\epsilon)$-approximated in time $2^{2^t}\tilde{O}(|X|+|Y|)$ with prob. 0.99, where

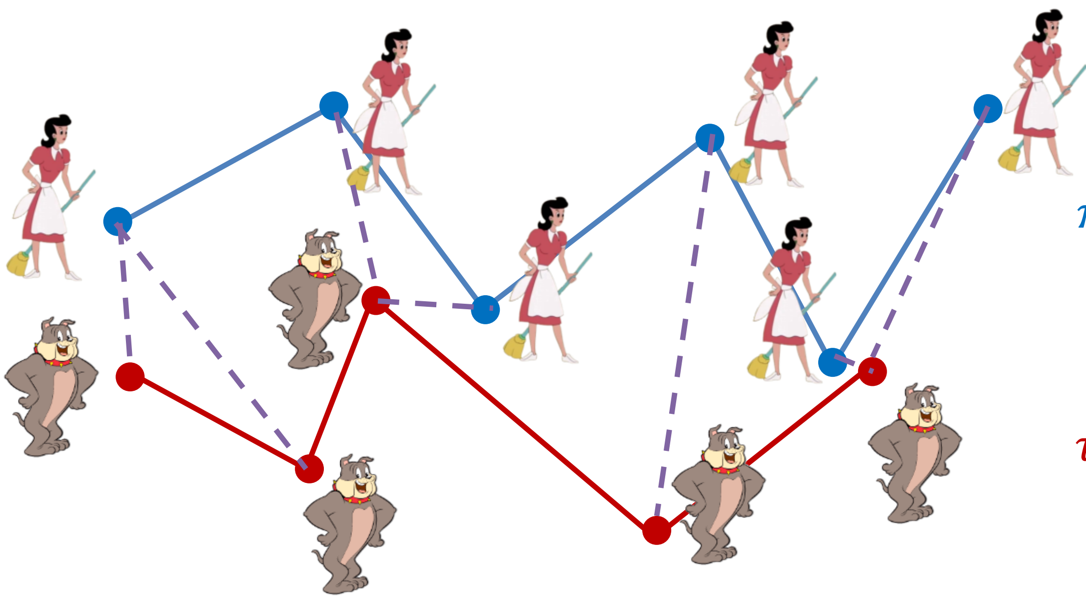
$$t = O(ddim \log(\mathrm{ddim}/\epsilon))$$



The underlying techniques are different from the previous theorem

Application: (k, ℓ)-median under discrete Frechet distance

Discrete Frechet distance: Measure the distance between polygonal curves



$\pi \in \mathbb{X}_6$

$\tau \in \mathbb{X}_5$

$\mathbb{X}_z$: Set of polygonal curves which

- Contain $\leq z$ vertices

$$\mathbf{d_{dF}}(\pi, \tau) \coloneqq \min_{\text{traversal } T} \max_{(i,j) \in T} \mathbf{d}\left(p_i, q_j\right)$$

Fact: $ddim(\mathbb{X}_z, \mathbf{d_{dF}}) = O(z)$

Application: (k,ℓ)-median under discrete Frechet distance

(k,ℓ)**-median under discrete Frechet distance** [Driemel, Krivosija, Sohler, SODA 2016] [Buchin, Driemel, Struijs, SWAT 2020]

Metric space $(X \cup Y, \mathbf{d})$

- $X \subset \mathbb{X}_z$, z is large, e,g., $z = \Omega(\log n)$
- $Y = \mathbb{X}_\ell$, $\ell = O(1)$
- $\mathbf{d}$ = discrete Frechet distance

Fact: $ddim(Y) = O(\ell)$

Corollary ((k,ℓ)-median)
Given a set of n polygonal curves $P \subset \mathbb{X}_z$, the (k,ℓ)-median can be $(1 + \epsilon)$-approximated in time $2^{2^t}\tilde{O}(\ell z n)$ with prob. 0.99, where

$$t = O(\ell \log(\ell/\epsilon))$$

First PTAS independent of k

- Previously: $\exp(k)$ dependence [Buchin, Driemel, Struijs, SWAT 2020] [Nath, Taylor, JoCG 2021]

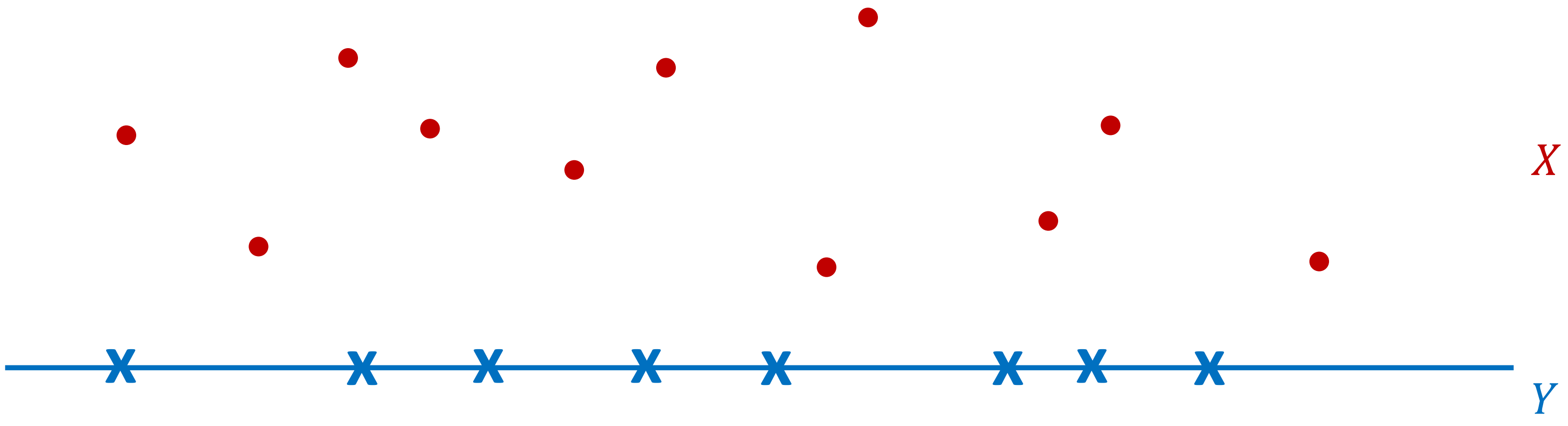
Overview of Techniques

with focus on the low-dim centers setting

Low-dimensional Centers

- Our setting:
  - High-dim clients: $ddim(X)$ unbounded
  - Low-dim centers: $ddim(Y) \le O(1)$

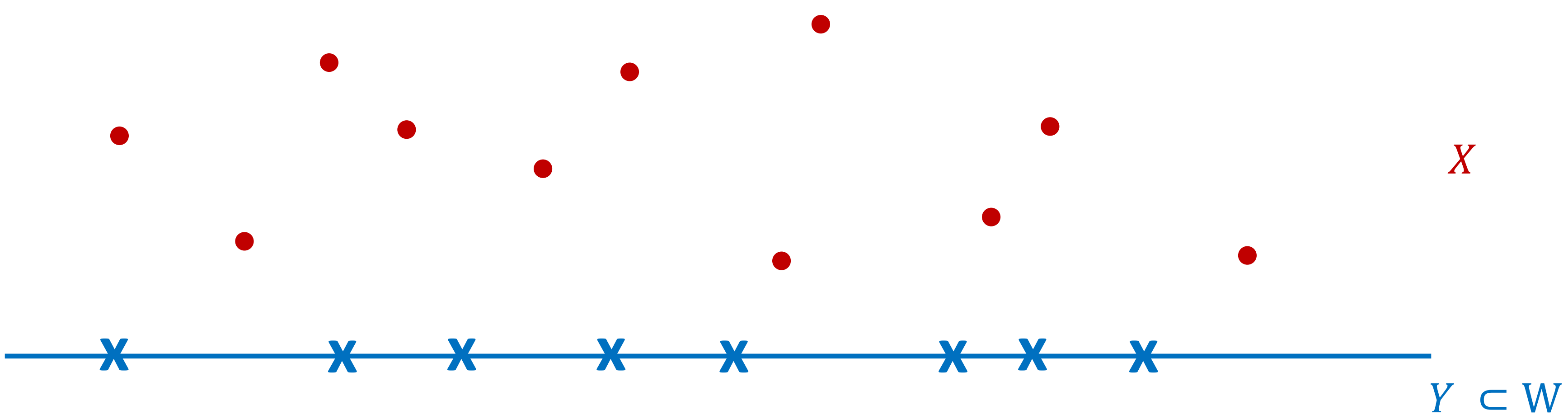
- Let's focus on the k-median problem.
  - $\min_{F \subseteq Y, |\mathrm{F}| = \mathrm{k}} \sum_{x \in X} \mathbf{d}(x, F)$



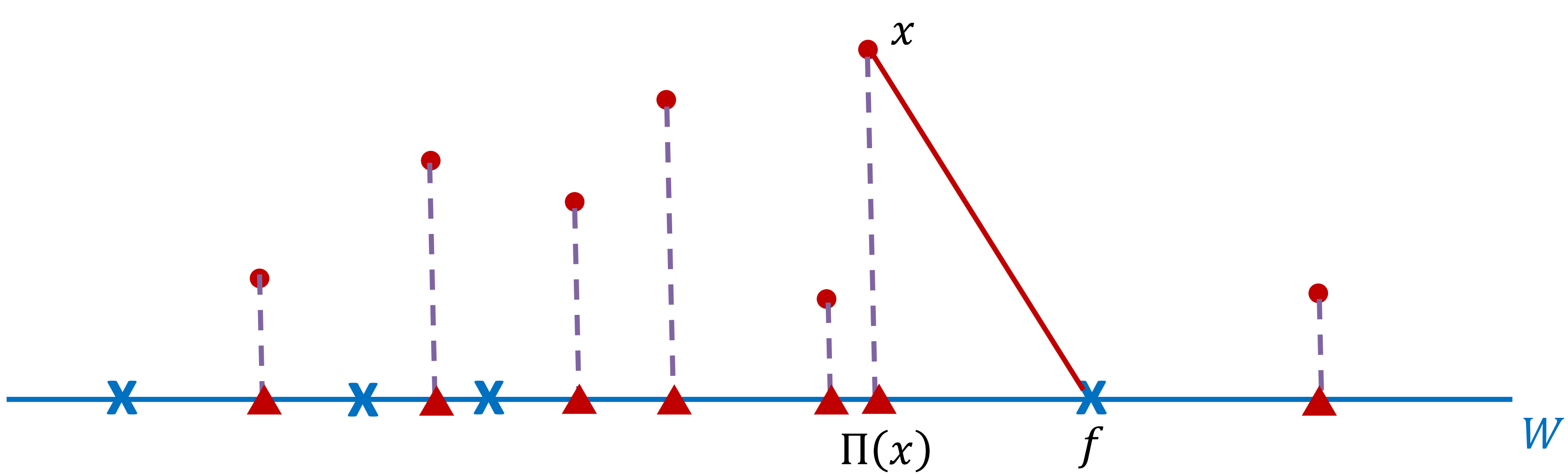
A Toy Problem

- Our setting:
  - High-dim clients: X = points in $\mathbb{R}^d$
  - Low-dim centers: $Y \subset O(1)$-dim subspace of $\mathbb{R}^d$

- k-means problem.
  - $\min_{F \subseteq Y, |F|=k} \sum_{x \in X} \mathbf{d}(x, F)^2$



Dimension Reduction



- Step 1: Π: Project every client to the low dimensional space Y
- Step 2: Compute $(1+\epsilon)$-apx solution F of $\Pi(X) \cup Y$
- Claim: F is a $(1+\epsilon)$-apx for $X \cup Y$

Proof: $\forall\, x \in X, \mathbf{d}(x, F)^2 = \boxed{\mathbf{d}(\Pi(x), F)^2} + \boxed{\mathbf{d}\big(x, \Pi(x)\big)^2}$ (Pythagorean!)

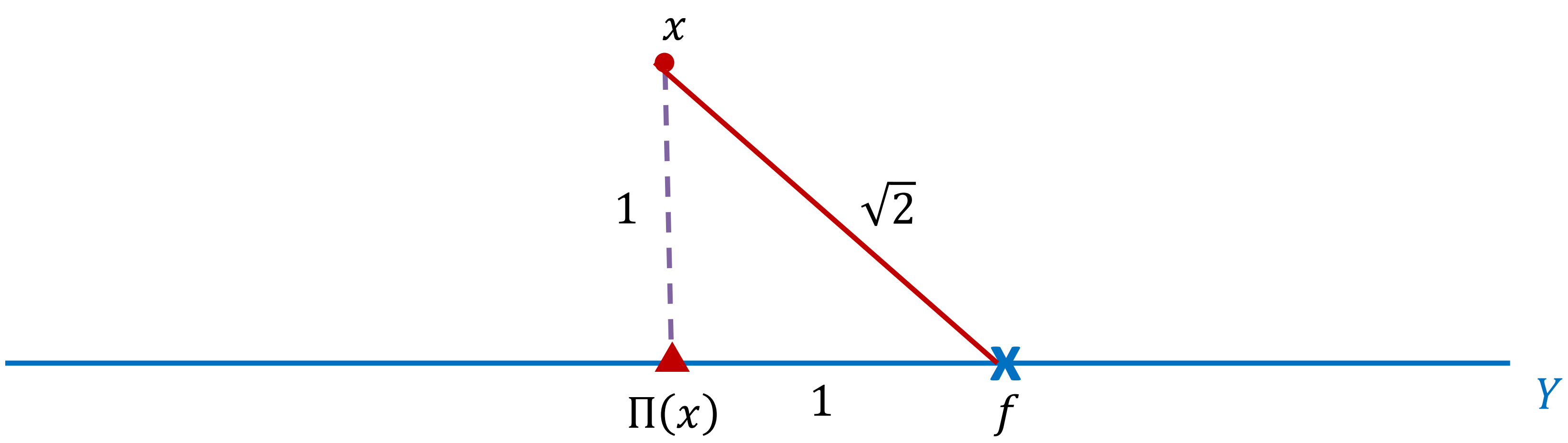
$(1+\epsilon)$-approximated

Fixed

Dimension Reduction

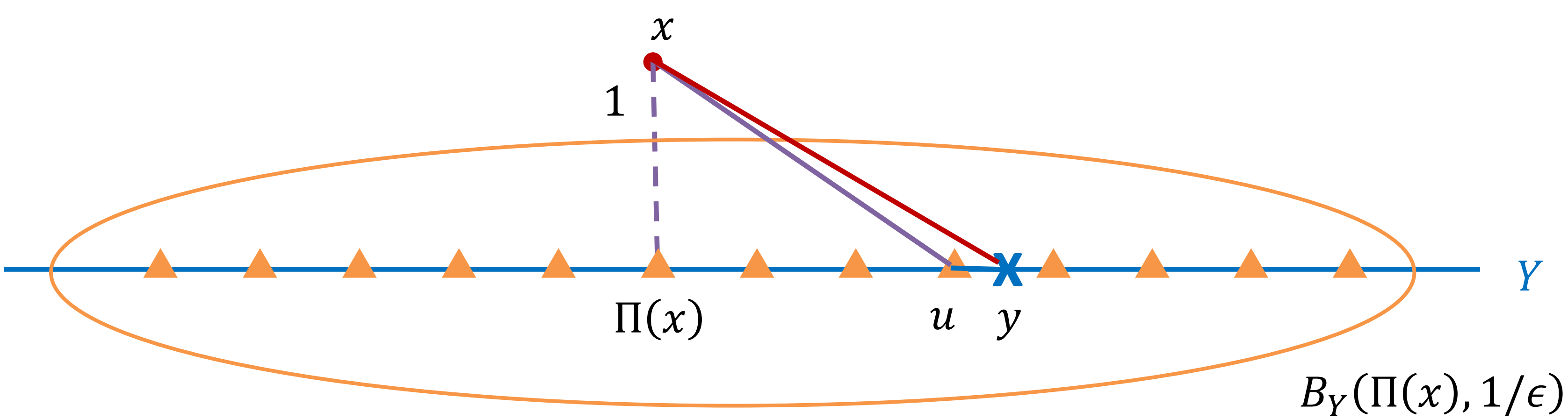
Issues:

- No Pythagorean in general metrics
- No Pythagorean for k-median



Proxy Set

- Idea: Instead of replacing x by a single point $\Pi(x) \in Y$, replace x by a set of points (called proxy set) $N_x \subseteq Y$
- Proxy set N_x is an ϵ-net of $B_Y(\Pi(x), 1/\epsilon)$



Lemma

Define $\mathbf{D}(x,y) \coloneqq \min_{u \in N_x} \mathbf{d}(x,u) + \mathbf{d}(u,y)$.

Then

$$\mathbf{d}(x,y) \leq \mathbf{D}(x,y) \leq (1+\epsilon)\mathbf{d}(x,y)$$

Dimension Reduction

- Proxy set N_x is an ϵ-net of $B_Y(\Pi(x), 1/\epsilon)$

Lemma

Define $\mathbf{D}(x, y) \coloneqq \min_{u \in N_x} \mathbf{d}(x, u) + \mathbf{d}(u, y)$.

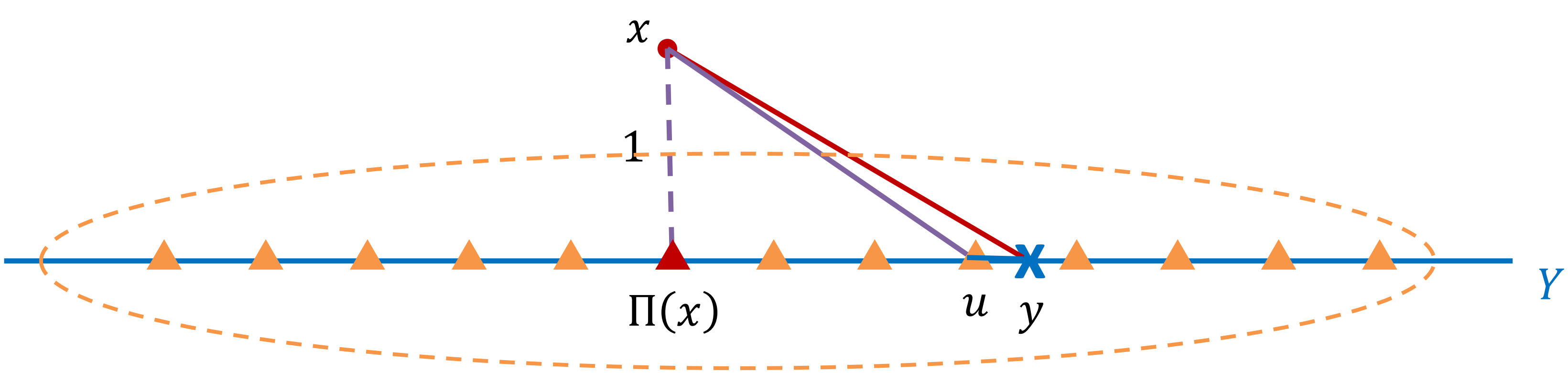
Then

$$\mathbf{d}(x, y) \leq \mathbf{D}(x, y) \leq (1 + \epsilon)\mathbf{d}(x, y)$$

Proof:

Case 1: $y \in B_Y(\Pi(x), 1/\epsilon)$

- $\exists u \in N_x$, s.t. $\mathbf{d}(u, y) \leq \epsilon$
- $\mathbf{D}(x, y) \leq \mathbf{d}(x, u) + \mathbf{d}(u, y) \leq (1 + \epsilon)\mathbf{d}(x, y)$



Dimension Reduction

- Proxy set N_x is an ϵ-net of $B_Y(\Pi(x), 1/\epsilon)$

Lemma

Define $\mathbf{D}(x,y) \coloneqq \min_{u \in N_x} \mathbf{d}\,(x,u) + \mathbf{d}(u,y)$.

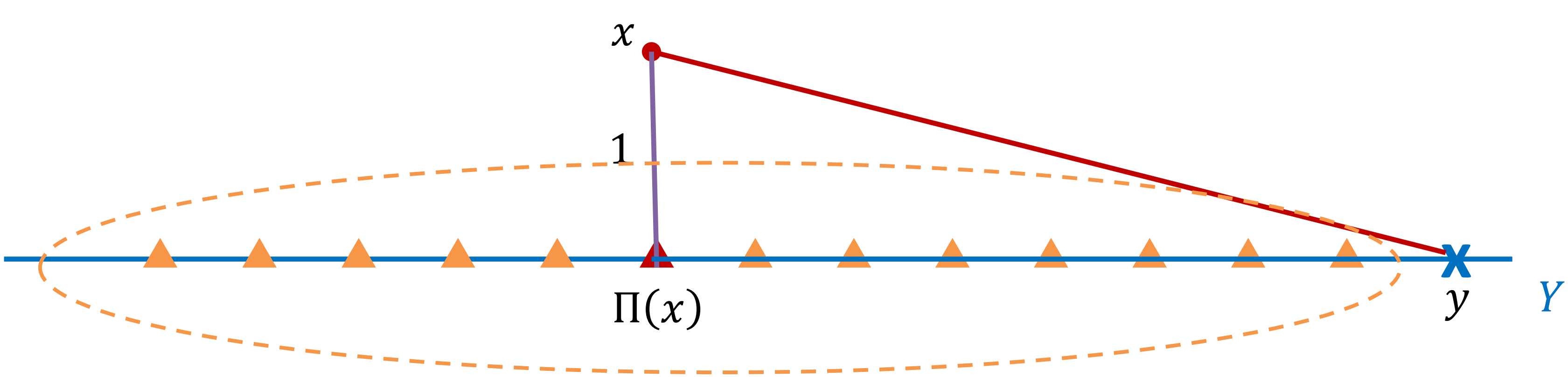
Then

$$\mathbf{d}(x,y) \le \mathbf{D}(x,y) \le (1+\epsilon)\mathbf{d}(x,y)$$

Proof:

Case 2: $y \notin B_Y(\Pi(x), 1/\epsilon)$

- Then $\mathbf{d}(\Pi(x), y) \ > \ 1/\epsilon$
- $\mathbf{D}(x,y) \le \mathbf{d}(x, \Pi(\mathrm{x})) + \mathbf{d}(\Pi(\mathrm{x}), y) \le (1+\epsilon)\mathbf{d}(x,y)$



Dimension Reduction

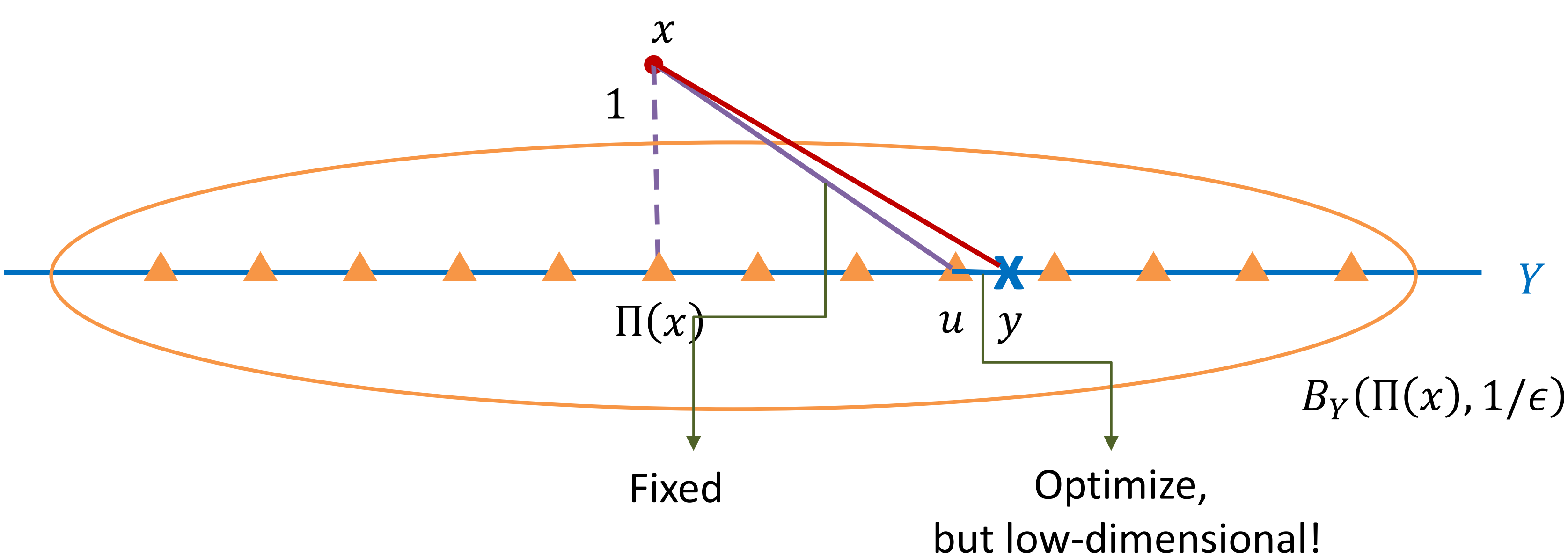
- Proxy set N_x is an ϵ-net of $B_Y(\Pi(x), 1/\epsilon)$

Lemma

Define $\mathbf{D}(x,y) := \min_{u \in N_x} \mathbf{d}(x,u) + \mathbf{d}(u,y)$.

Then

$$\mathbf{d}(x,y) \le \mathbf{D}(x,y) \le (1+\epsilon)\mathbf{d}(x,y)$$



Next Steps

Integrate our dimension reduction with the low-dim PTAS [Cohen-Addad, Feldmann, Saulpic, J. ACM 2021]

- Non-blackbox
- We do it via a careful modification of [Cohen-Addad, Feldmann, Saulpic, J. ACM 2021]
- Technical challenge: Maintain the connection cost for every x

Thanks for Your Attention!