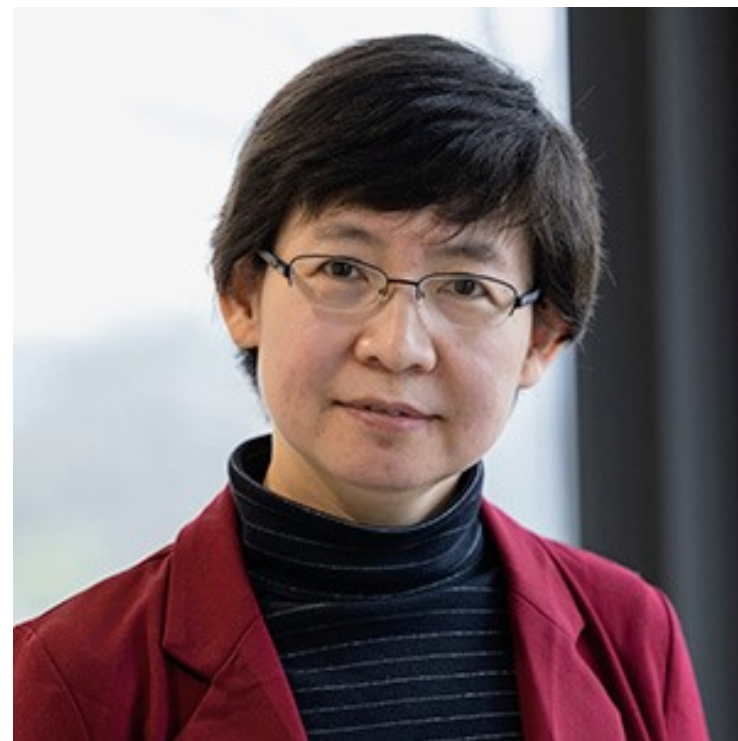


Locality Sensitive Hashing in Hyperbolic Space

Chengyuan Deng
Rutgers University

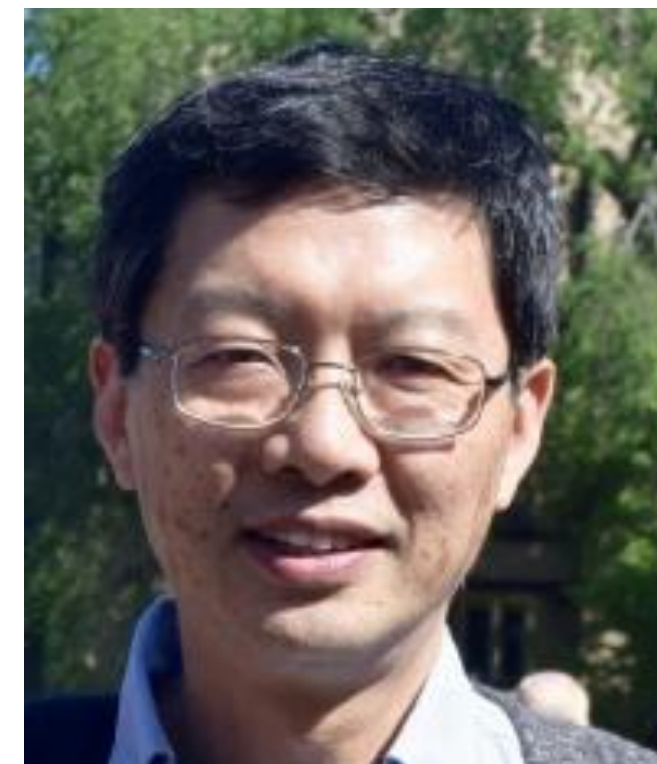
Joint work with



Jie Gao



Kevin Lu



Feng Luo

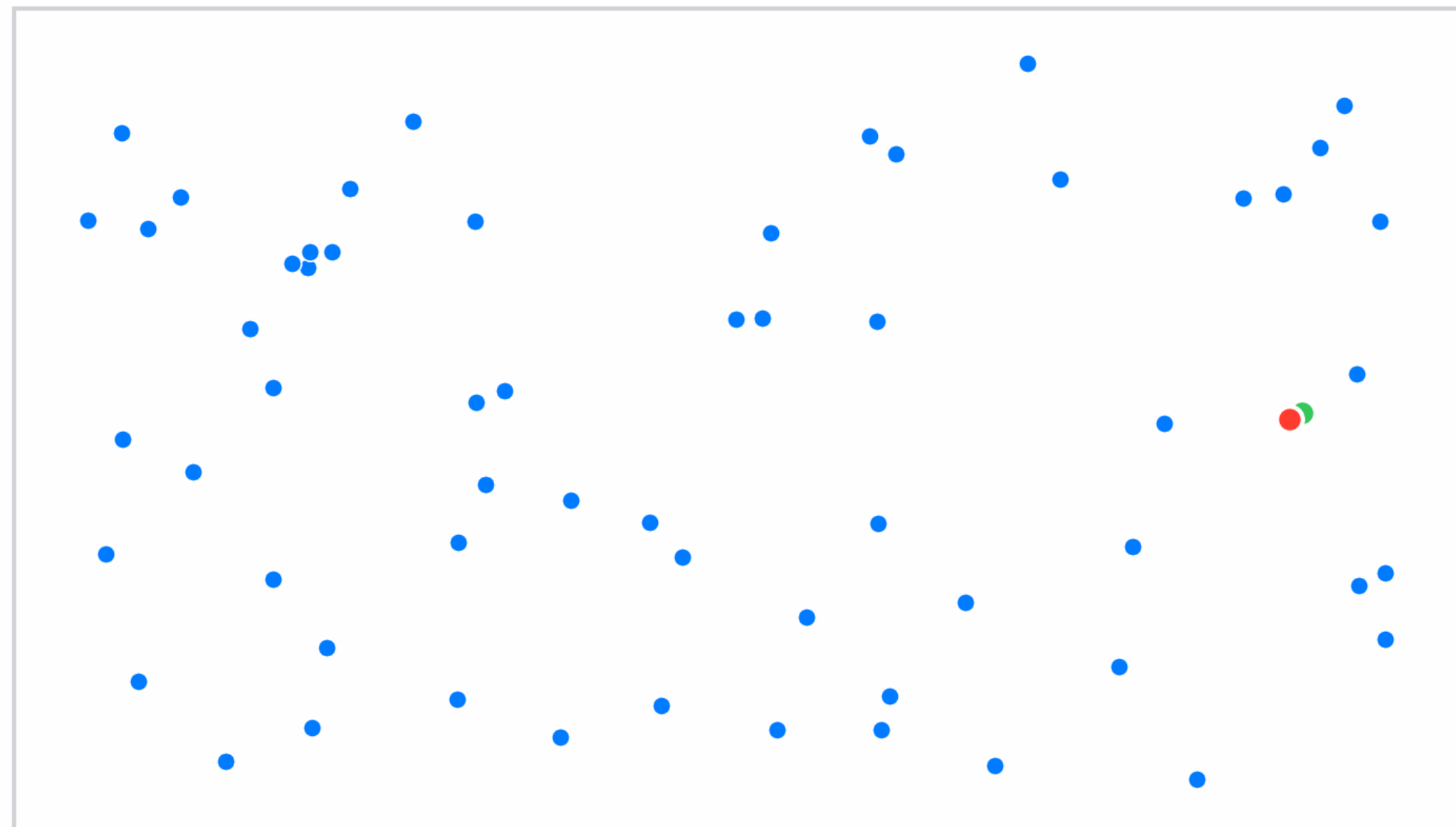


Cheng Xin

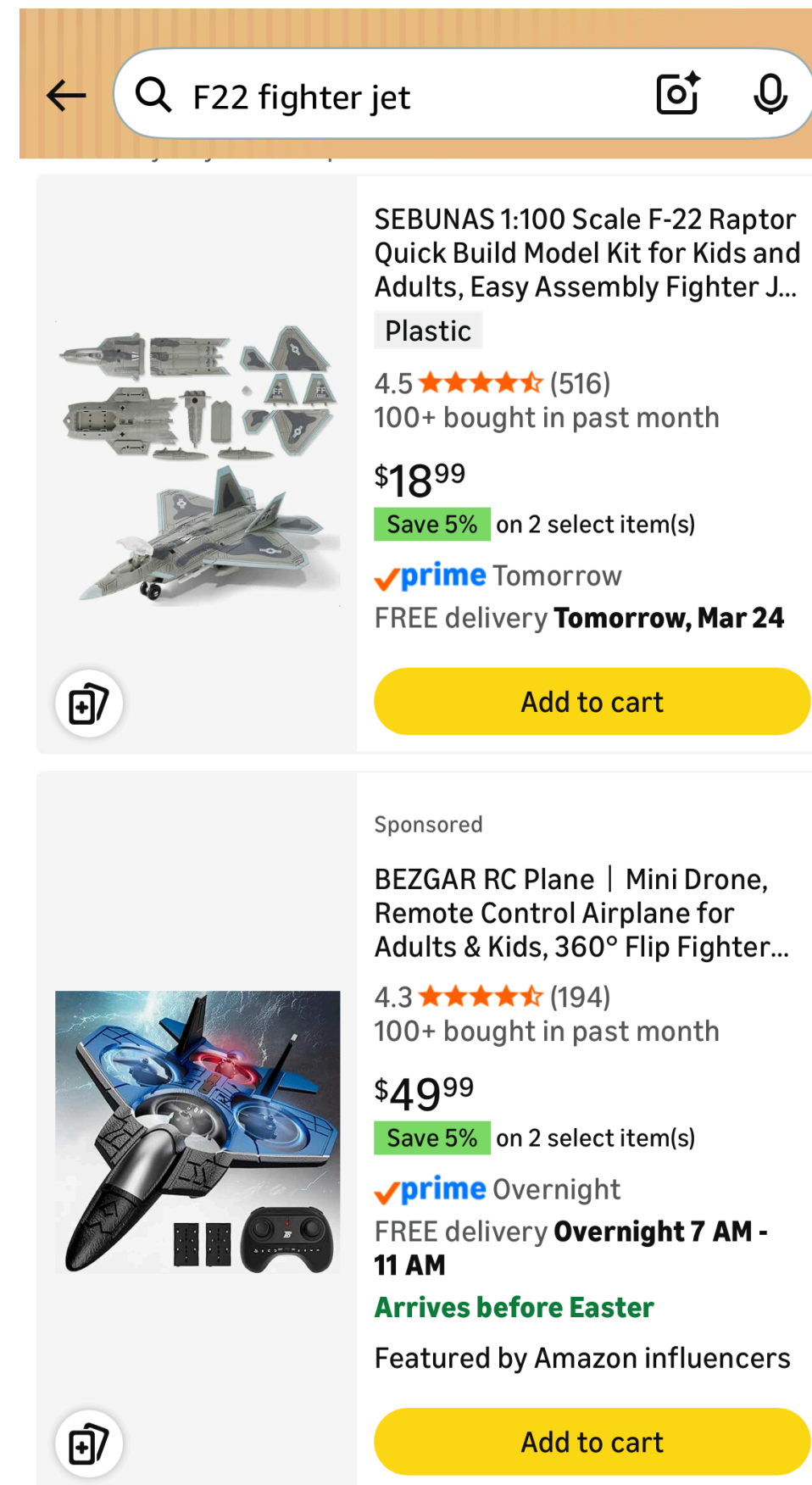
Nearest Neighbor Search

Given a metric space $M = (\mathcal{X}, d)$, a point set $P \subset \mathcal{X}$ of size n , a query input $q \in \mathcal{X}$, find $p \in P$ such that

$$d(p, q) = \min_{x \in P} d(x, q)$$

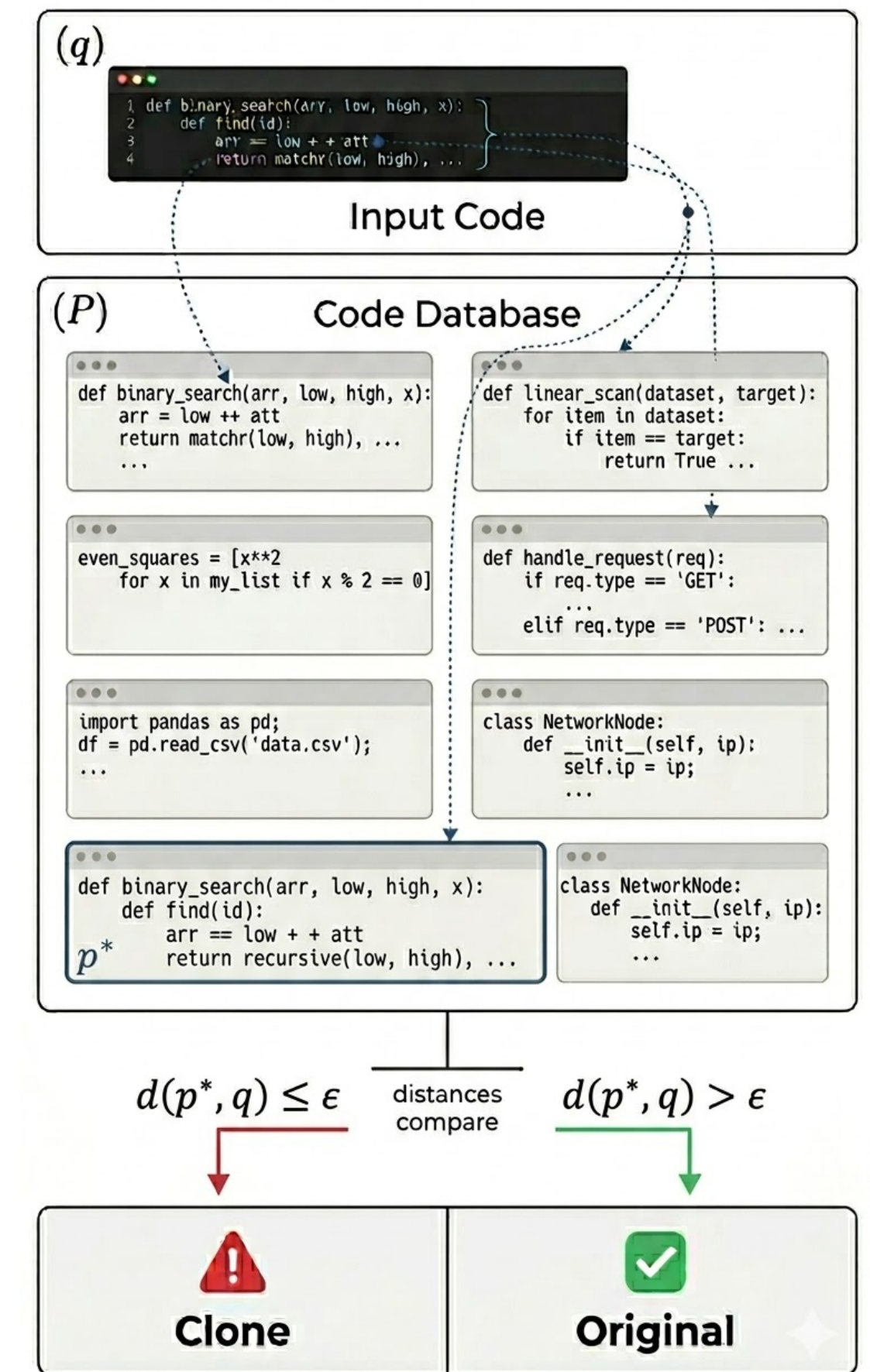


Nearest Neighbor Search



- Listen to a music snippet and guess the song

- E-commerce item recommendation



- (GenAI) Code clone and plagiarism detection

Nearest Neighbor Search

- Goal: fast query time and small storage for the database
- Works in low-dimension, e.g. KD tree
- Curse of dimensionality: all known algorithms degrade rapidly with d

Algorithm	Query Time	Space
Full Indexing	$O(d \log n)$	$n^{O(d)}$ (Voronoi diagram)
Without Indexing (Linear Scan)	$O(dn)$	$O(dn)$

Approximate Nearest Neighbor Search

Given a metric space $M = (\mathcal{X}, d)$, a point set $P \subset \mathcal{X}$ of size n , a query input $q \in \mathcal{X}$, find $p \in P$ such that

$$d(p, q) \leq c \cdot \min_{x \in P} d(x, q)$$

Vast literature:

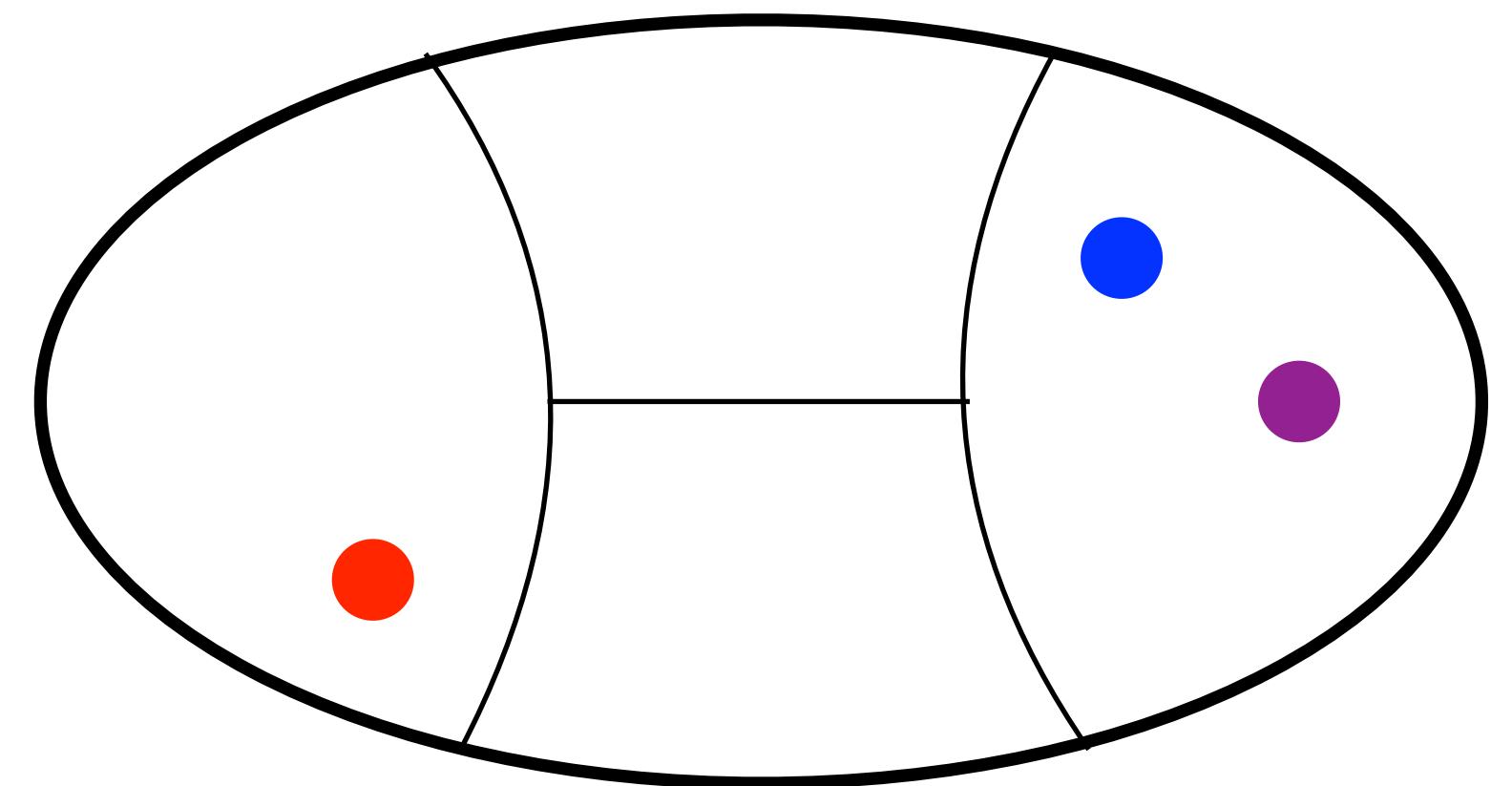
- Milder dependence on dimension
[Arya-Mount '93], [Clarkson '94], [Arya-Mount-Netanyahu-Silverman-We '98], [Kleinberg '97], [Har-Peled '02], ...
- Little to no dependence on dimension
 - **Dimension reduction:** [Kushilevitz-Ostrovsky-Rabani '98], [Ailon-Chazelle '06]
 - **LSH:** [Indyk-Motwani '98], [Gionis-Indyk-Motwani '99], [Charikar '02], [Datar-Immorlica-Indyk-Mirroknii '04], [Andoni-Indyk '06], ...
 - **Navigable graph:** [Laarhoven '18], [Prokhorenkova-Shekhovtsov '20], [Indyk-Xu '23], [Musco-Silwal-Xu '25], ...

Locality Sensitive Hashing

Given a point set P in a metric space $(\mathcal{X}, d)$, LSH maps each point into a bucket

1. Close points are likely in the same bucket

2. Far points are likely in different buckets



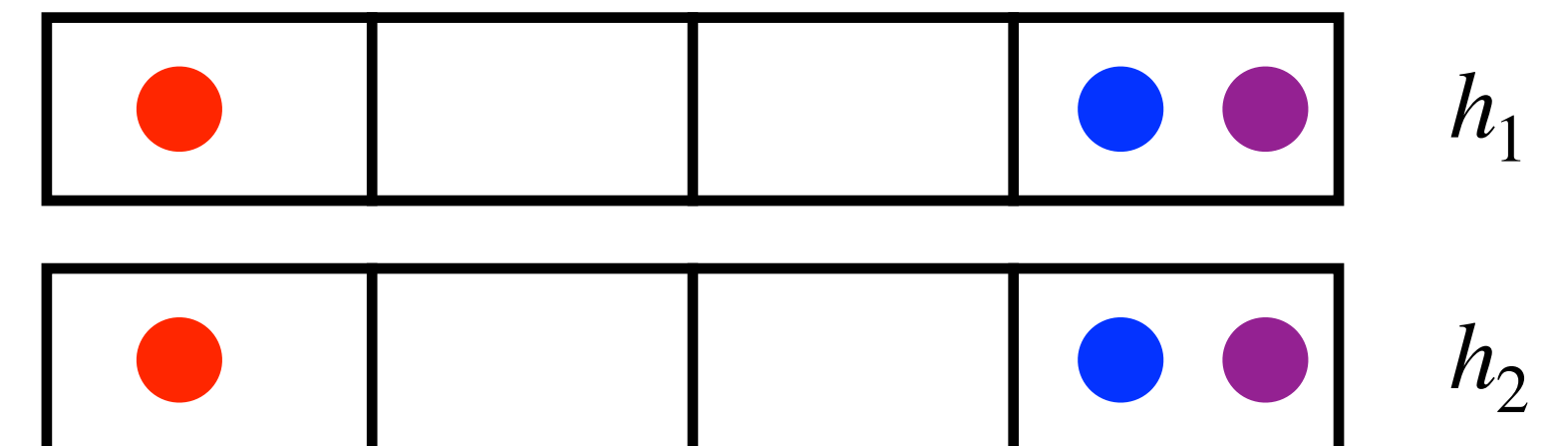
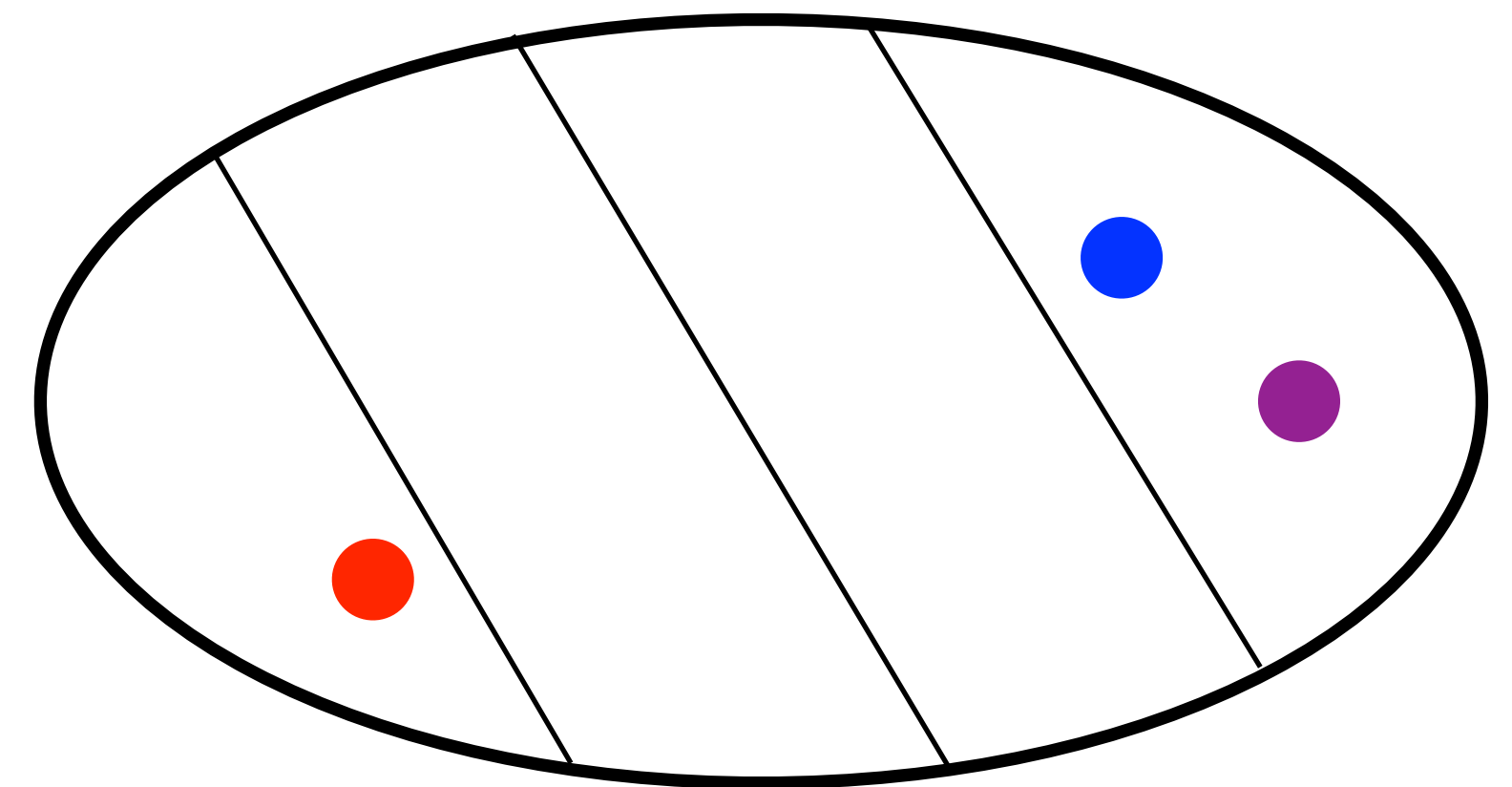
Locality Sensitive Hashing

Given a point set P in a metric space $(\mathcal{X}, d)$, LSH maps each point into a bucket

1. Close points are likely in the same bucket

2. Far points are likely in different buckets

• LSH is essentially a space partition scheme



Locality Sensitive Hashing

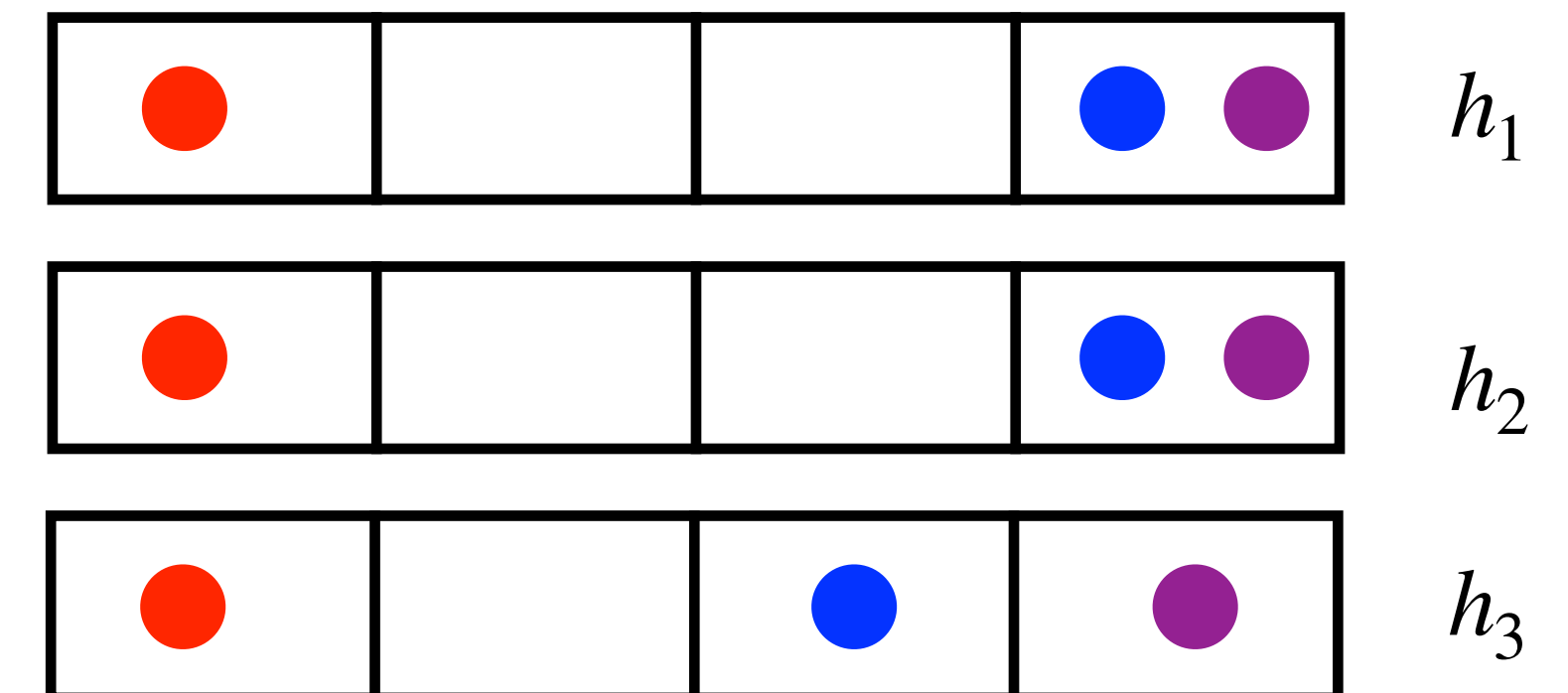
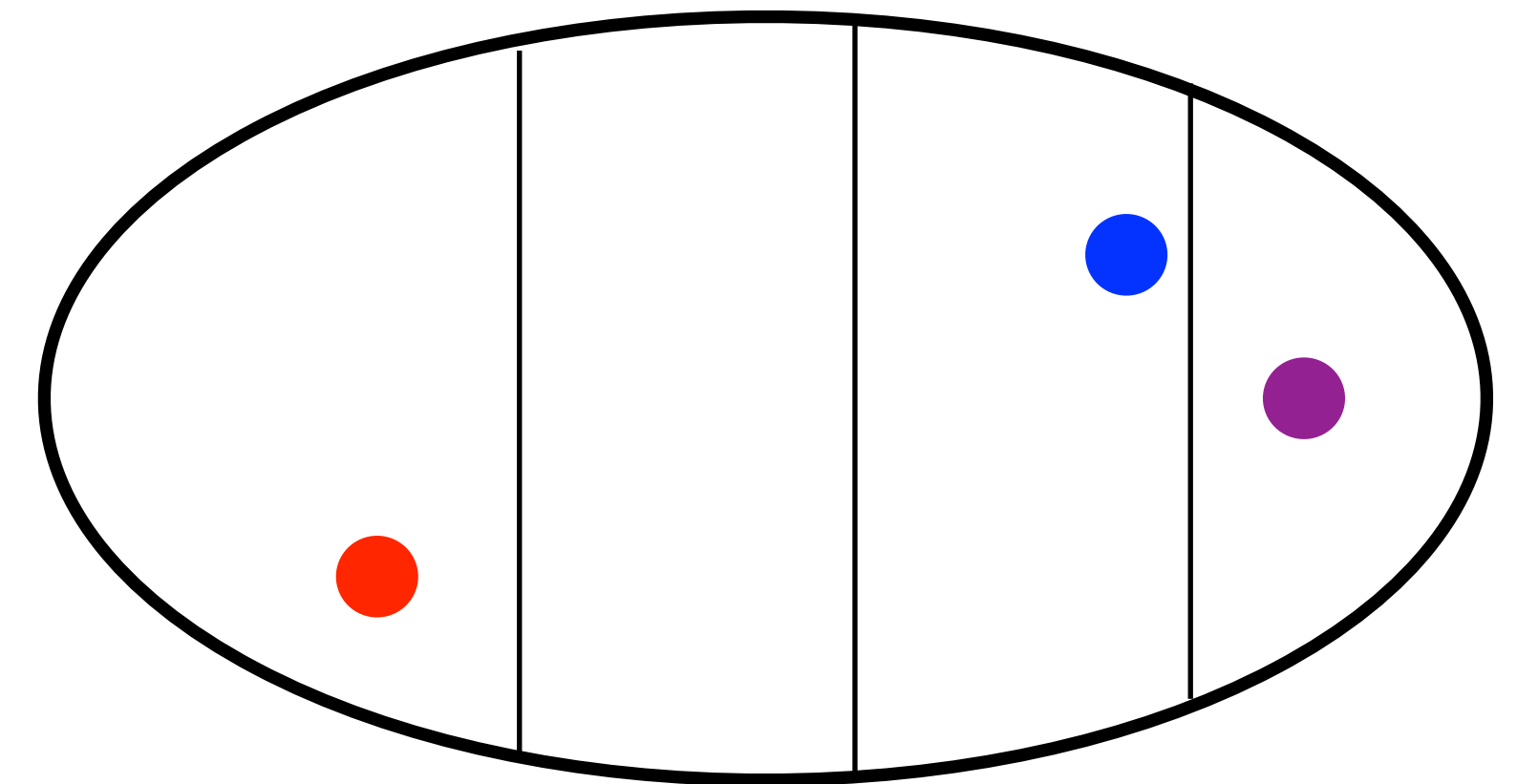
Given a point set P in a metric space $(\mathcal{X}, d)$, LSH maps each point into a bucket

1. Close points are likely in the same bucket

2. Far points are likely in different buckets



• LSH is essentially a random space partition scheme



Locality Sensitive Hashing

Given a point set P in a metric space $(\mathcal{X}, d)$, LSH maps each point into a bucket

1. Close points are likely in the same bucket

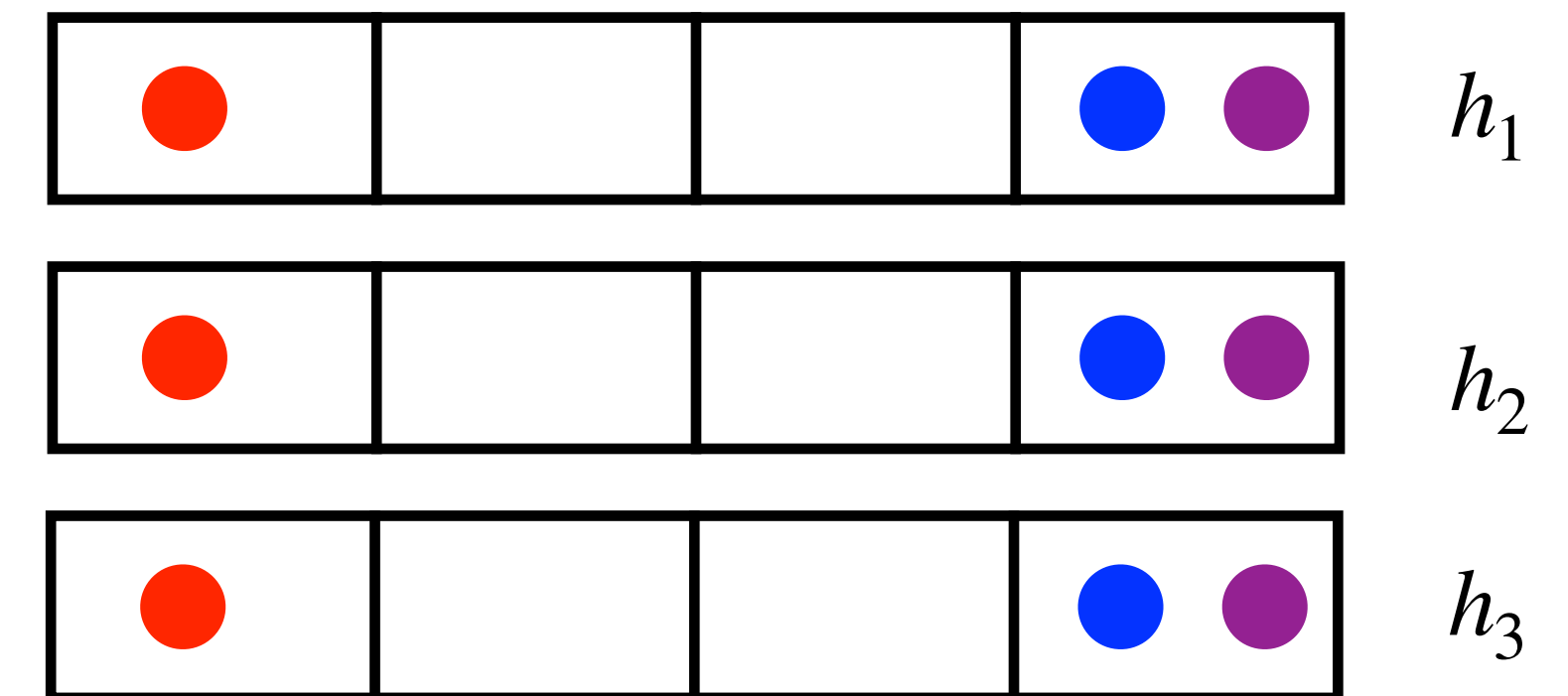
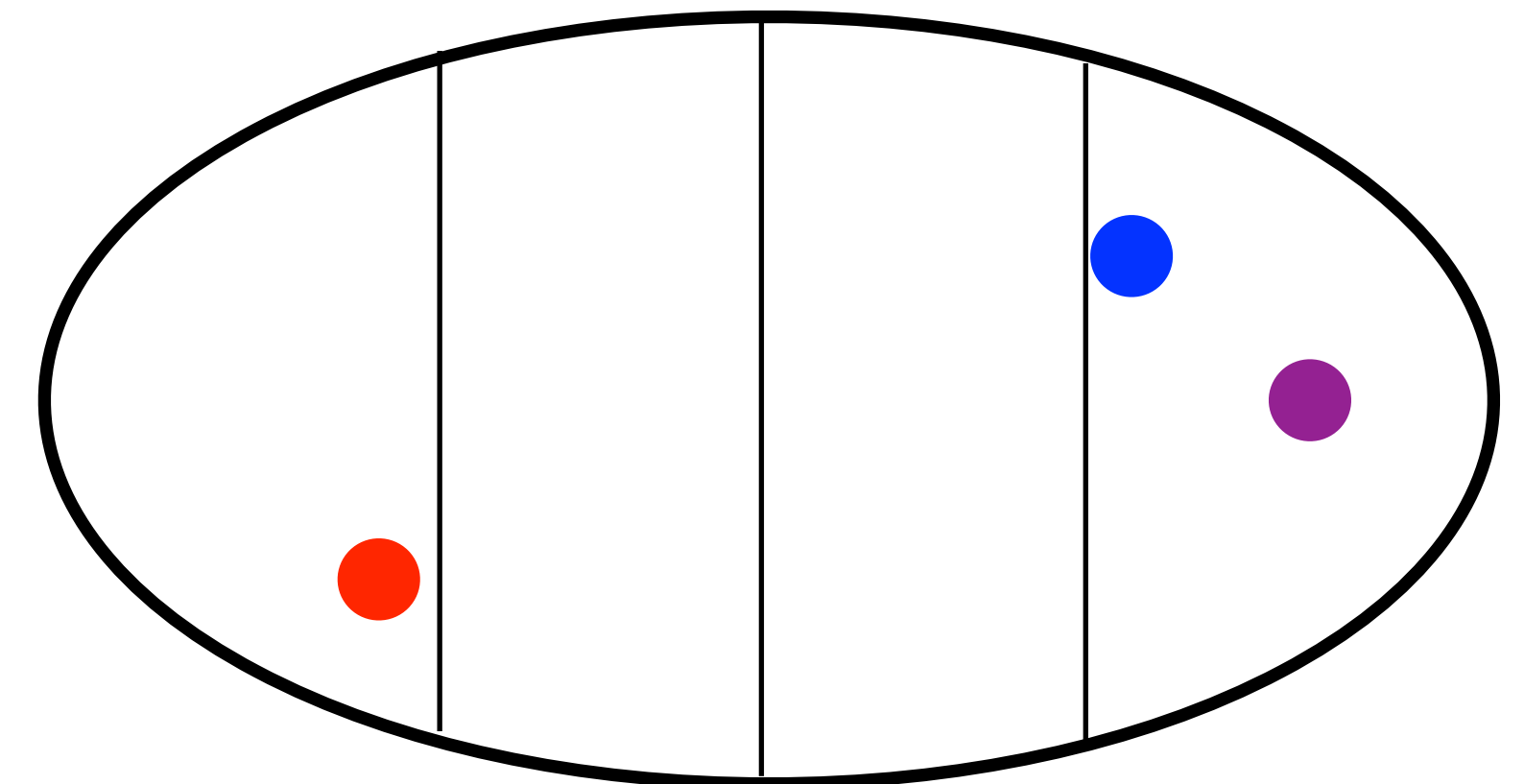
2. Far points are likely in different buckets



- LSH is essentially a random space partition scheme

- Repetition of hashing helps

👉 We are looking for a locality sensitive hash family



...

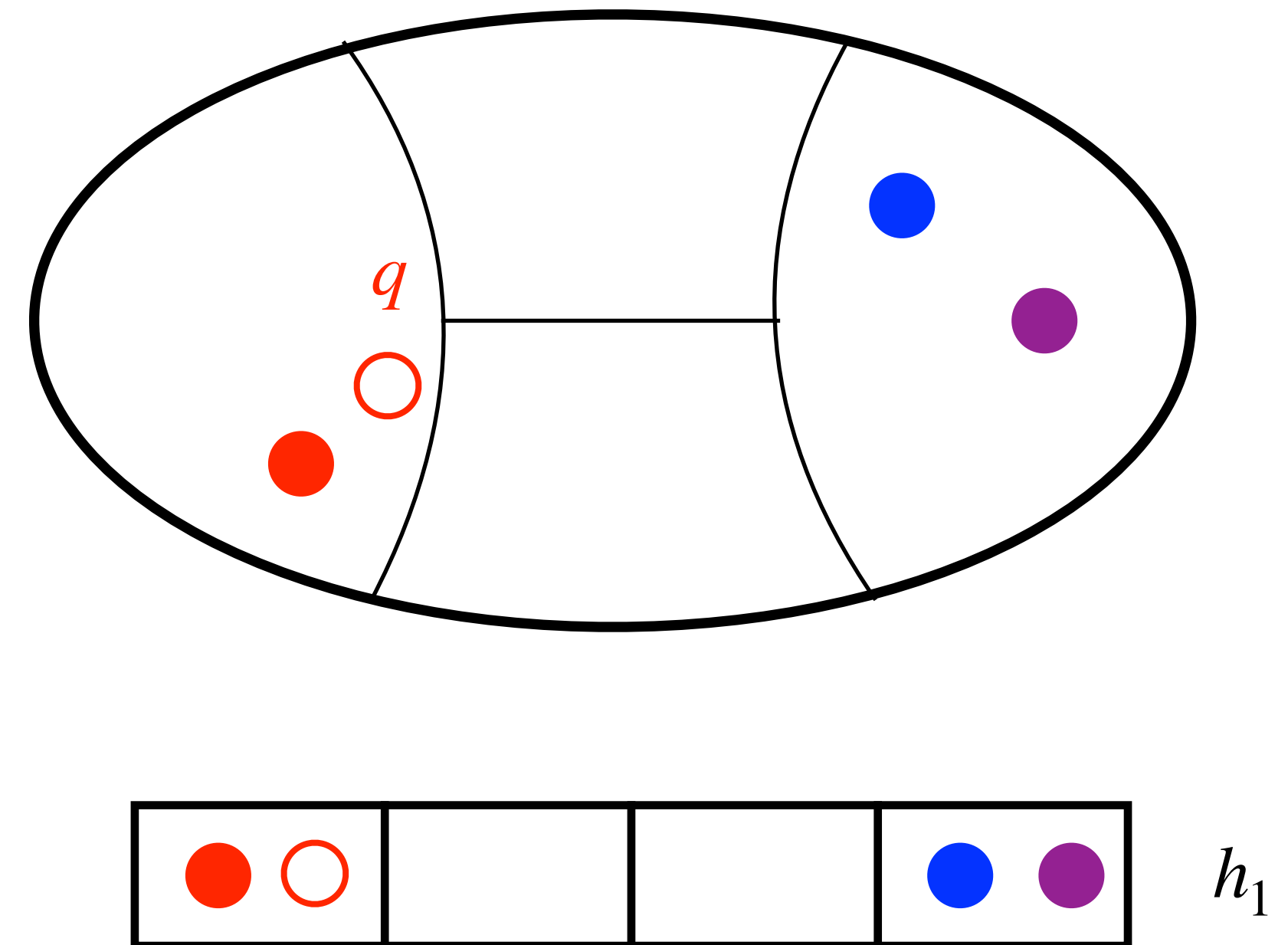
Locality Sensitive Hashing

Given a metric space $(\mathcal{X}, d)$, let U be a countable set and $H : \mathcal{X} \rightarrow U$ be a family of functions. We say H is a (r, cr, p_1, p_2) -sensitive family, where $r > 0$, $c > 1$, $0 < p_2 < p_1 < 1$, such that for any $x, y \in \mathcal{X}$:

1. If $d(x, y) \leq r$, $\Pr_{h \sim H} [h(x) = h(y)] \geq p_1$
2. If $d(x, y) \geq cr$, $\Pr_{h \sim H} [h(x) = h(y)] \leq p_2$

🔑 LSH is a good candidate to solve ANNS

Suppose $p_1 \approx 1, p_2 \approx 0$



Locality Sensitive Hashing

Given a metric space $(\mathcal{X}, d)$, let U be a countable set and $H : \mathcal{X} \rightarrow U$ be a family of functions. We say H is a (r, cr, p_1, p_2) -sensitive family, where $r > 0$, $c > 1$, $0 < p_2 < p_1 < 1$, such that for any $x, y \in \mathcal{X}$:

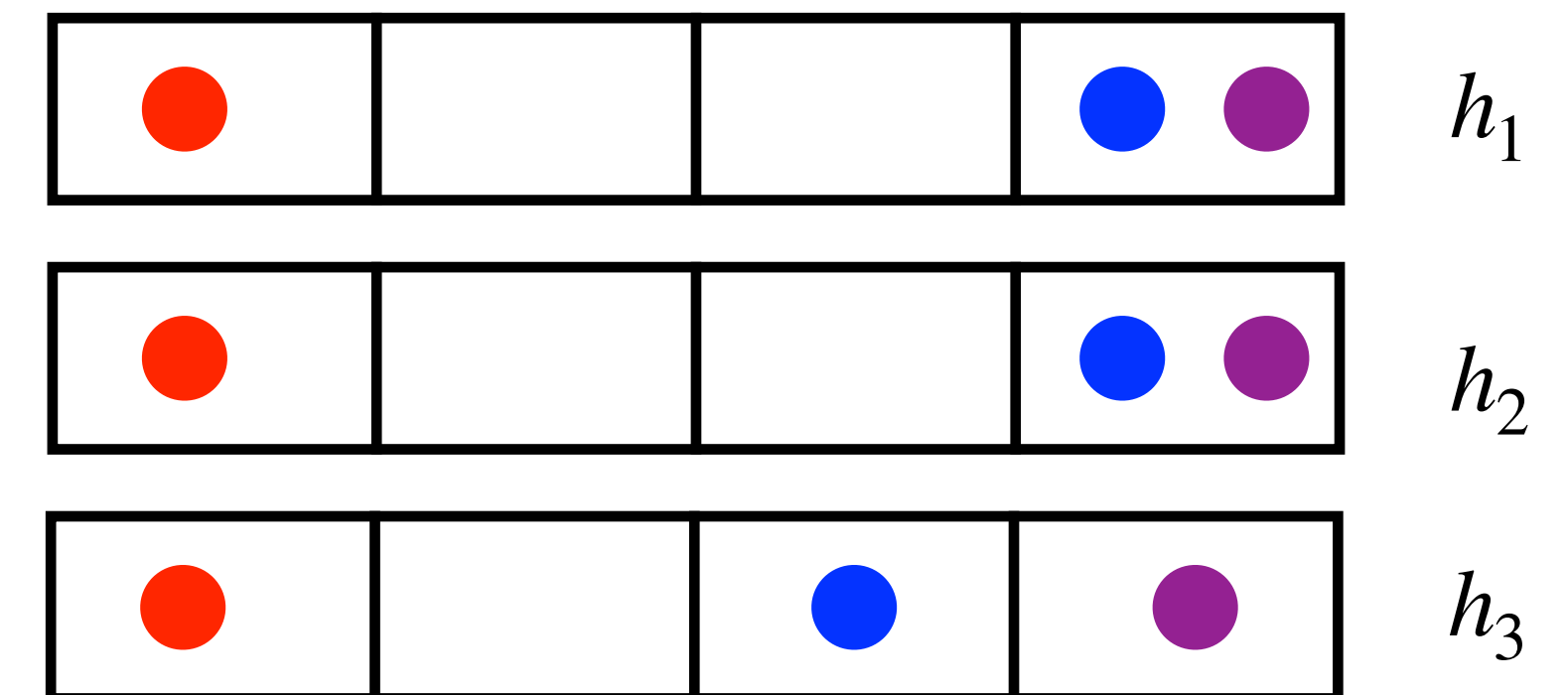
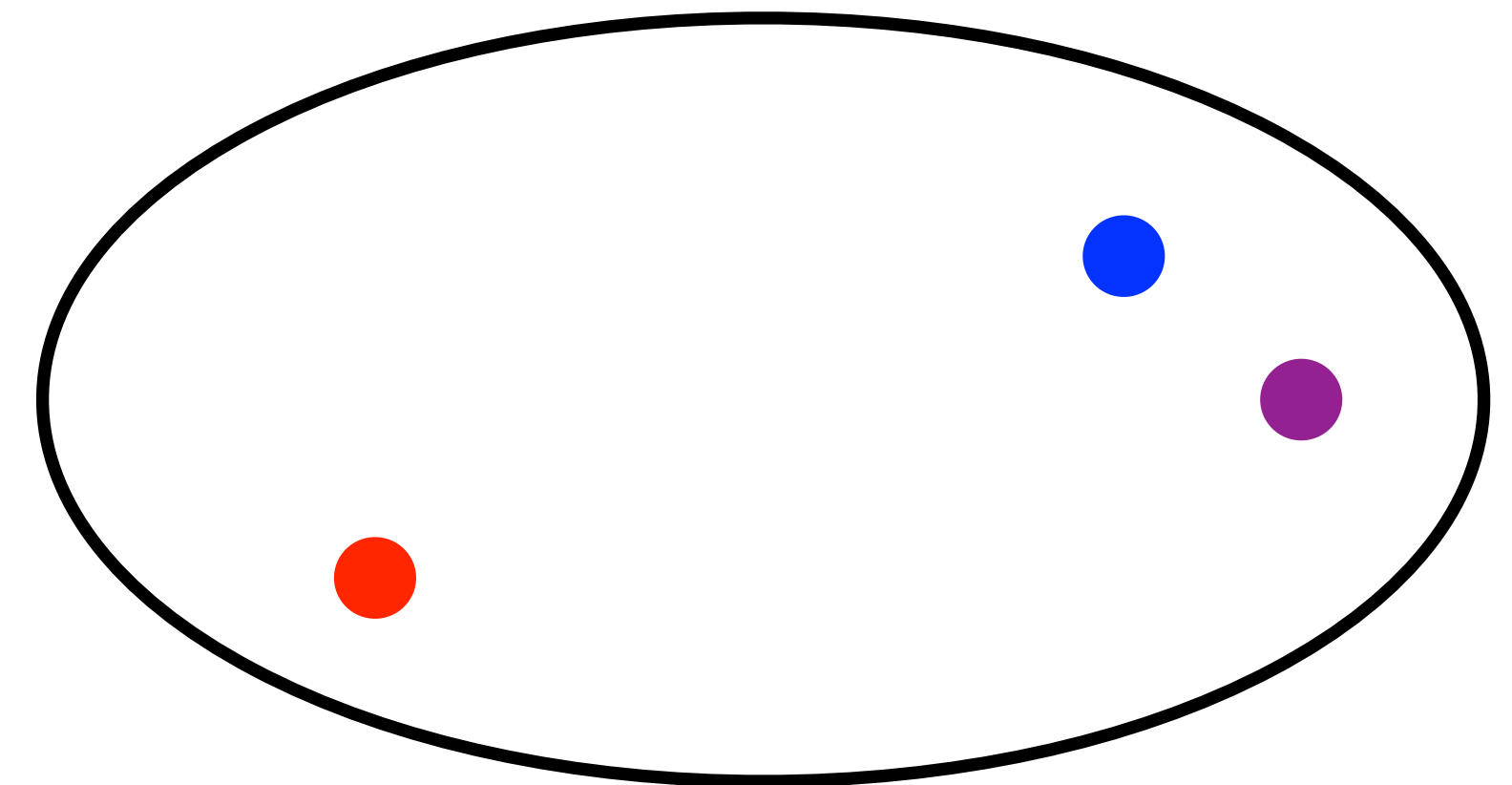
1. If $d(x, y) \leq r$, $\Pr_{h \sim H} [h(x) = h(y)] \geq p_1$
2. If $d(x, y) \geq cr$, $\Pr_{h \sim H} [h(x) = h(y)] \leq p_2$

🔍 How many hash tables do we need?

n^ρ , where $\rho = \log p_1 / \log p_2$

📌 Query time: $O(d \cdot n^\rho)$, Space: $O(n^{1+\rho} + dn)$

ρ is regarded as a quality measure for LSH, as a function of c



...

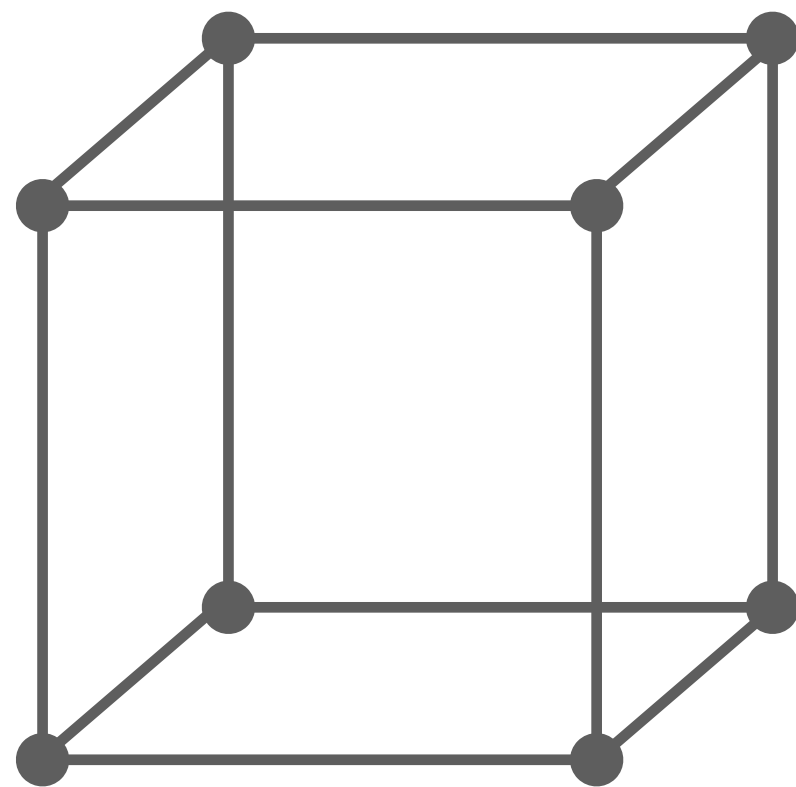
Locality Sensitive Hashing: Existing results

- LSH (later) finds more applications:

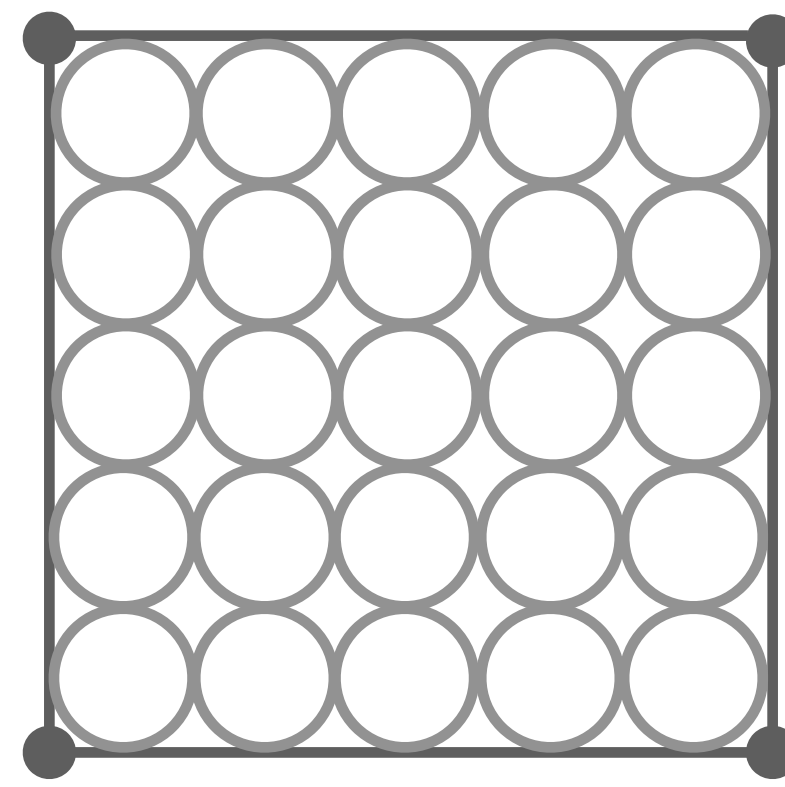
Bichromatic closest pair, Clustering, Kernel density estimation, Maximum inner product search...

- LSH (for today) is data-independent: only use **geometry of the metric space**

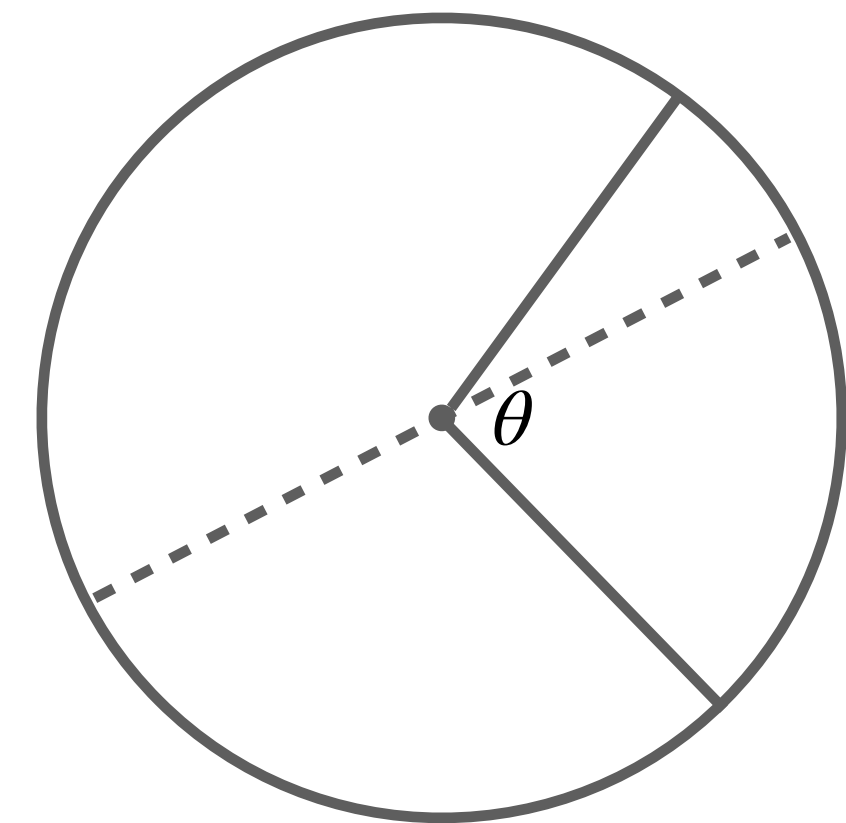
What metric spaces have been studied?



- **Hamming space** $\rho \approx 1/c$
Random coordinate sampling



- **Euclidean space** $\rho \approx 1/c^2$
Ball packing with random shift



- **Angular distance** $\rho \approx 1/c^2$
Random hyperplane rounding

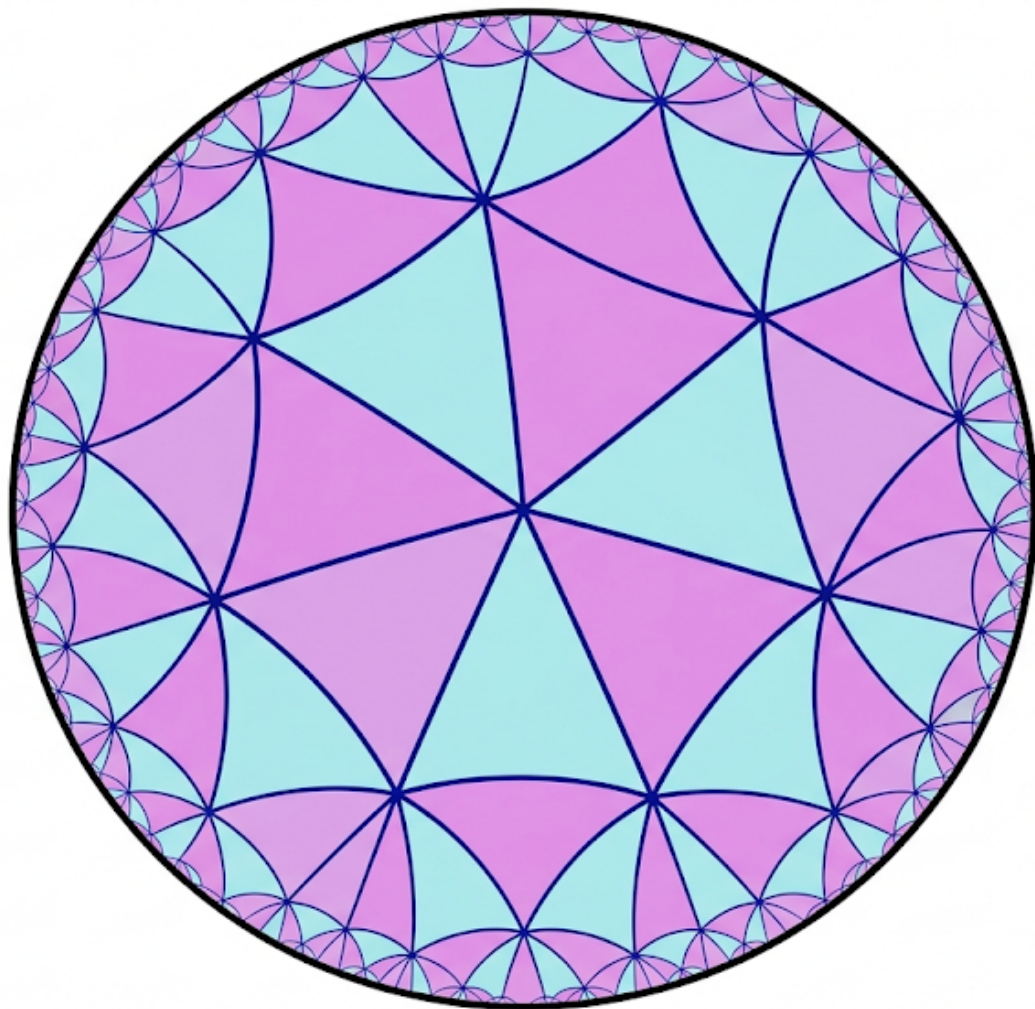
Locality Sensitive Hashing: Our results

- LSH (later) finds more applications:

Bichromatic closest pair, Clustering, Kernel density estimation, Maximum inner product search...

- LSH (for today) is data-independent: only use **geometry of the metric space**

What metric spaces have been studied?



This talk: LSH in Hyperbolic Space

	$d = 2$	$d \geq 3$
Upper bound	$\rho \leq 1/c$	$\rho \leq 1.59/c$
Lower bound	$\rho \geq 1/c^2 - o(1)$	

Hyperbolic Geometry

Hyperbolic space $\mathbb{H}^d$

simply connected, negatively curved d -dimensional Riemannian manifold

“ We understand new concepts by relating them to familiar ones.”

— Henri Poincaré

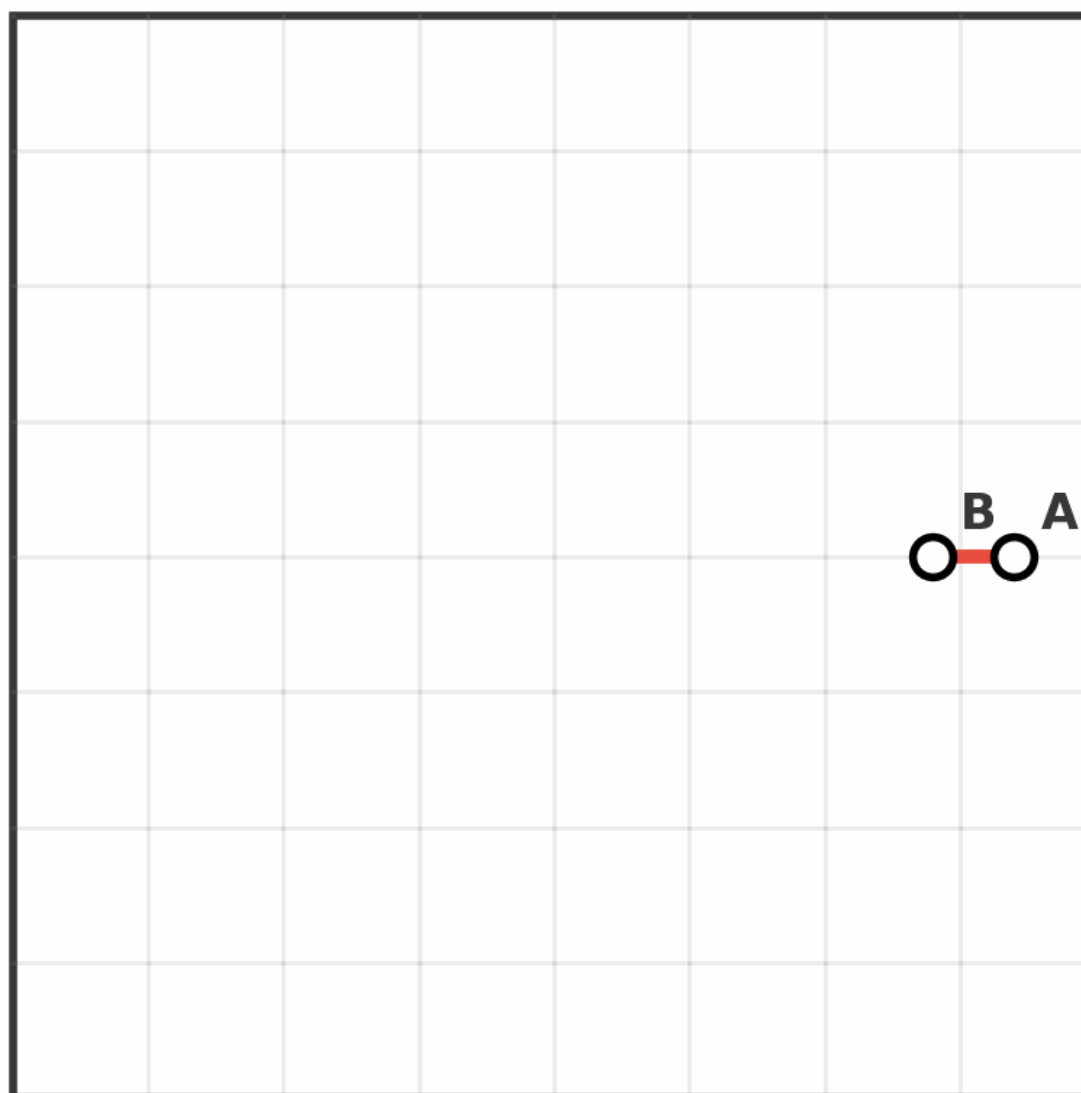
Hyperbolic Geometry

Hyperbolic space $\mathbb{H}^d$

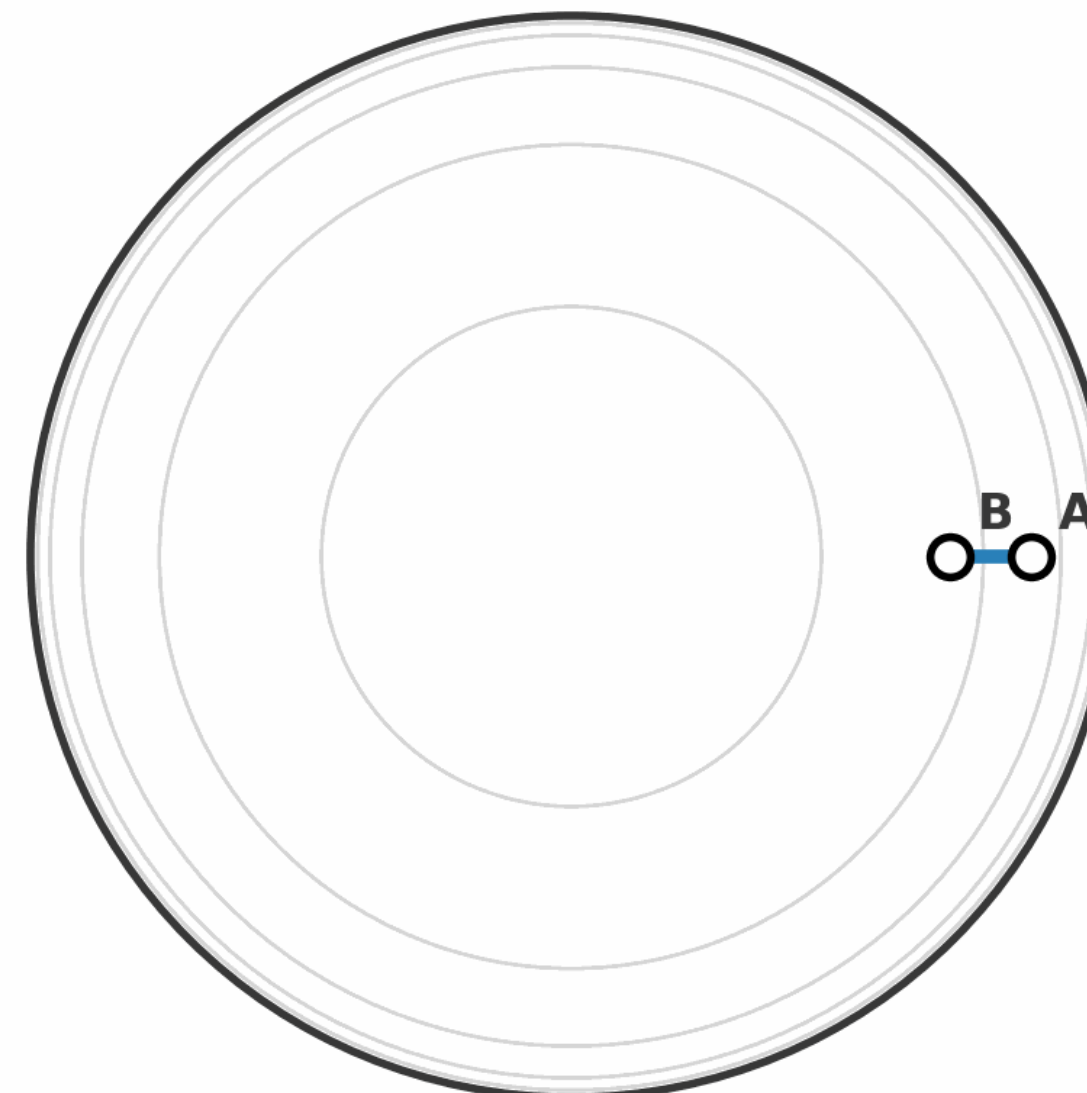
simply connected, negatively curved d -dimensional Riemannian manifold

Hyperbolic VS Euclidean: [geodesics](#)

Euclidean Space ($\mathbb{E}^2$)



Poincaré Disk ($\mathbb{D}^2$)



Rings show constant Hyperbolic step.
Boundary = Infinity ($d \rightarrow \infty$)

Live Telemetry

Point A (x, y):
[+0.850, +0.000]

Point B (x, y):
[+0.700, -0.000]

Euclidean Distance:
0.150

Hyperbolic Distance:
0.778

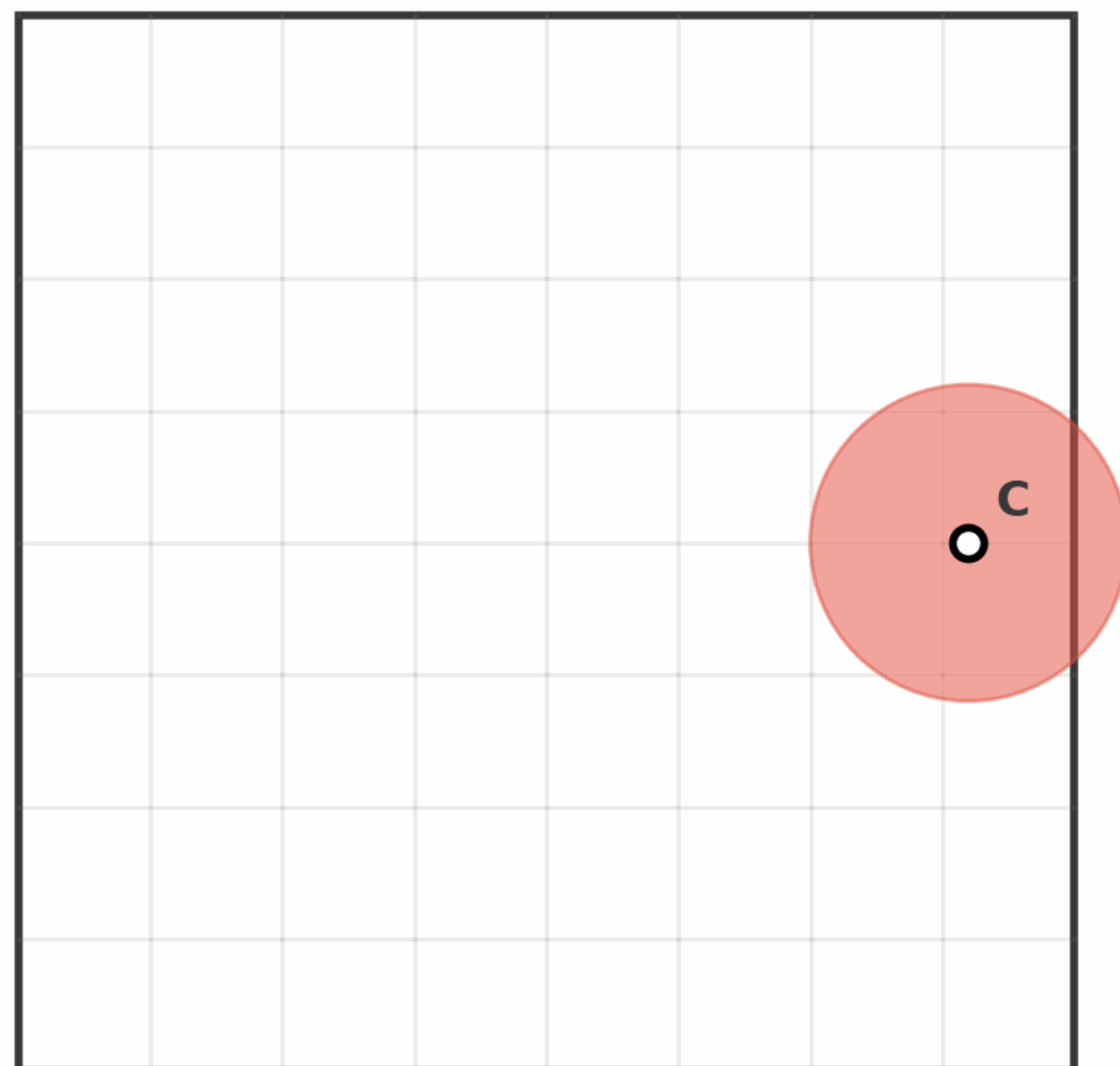
Hyperbolic Geometry

Hyperbolic space $\mathbb{H}^d$

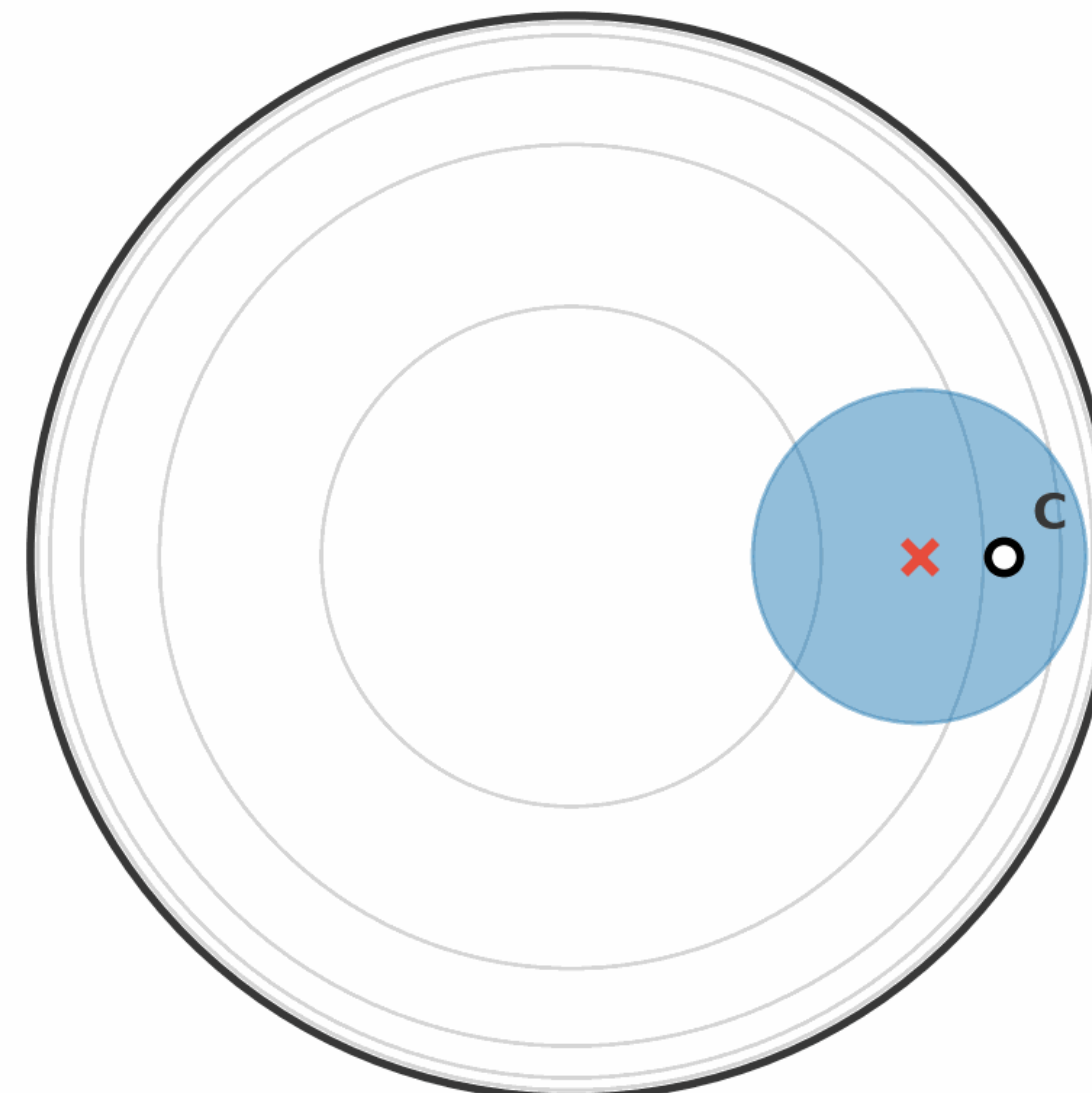
simply connected, negatively curved d -dimensional Riemannian manifold

Hyperbolic VS Euclidean: **balls**

Euclidean Ball ($\mathbb{E}^2$)



Hyperbolic Ball in $\mathbb{D}^2$



Constant Hyperbolic Radius ($R_H = 1.5$)

Live Telemetry

Center 'C' (x, y):
[+0.800, +0.000]

Euclidean Offset:
0.157

Hyperbolic Radius:
1.500 (Constant)

Apparent Euclidean:
0.308 (Shrinking)

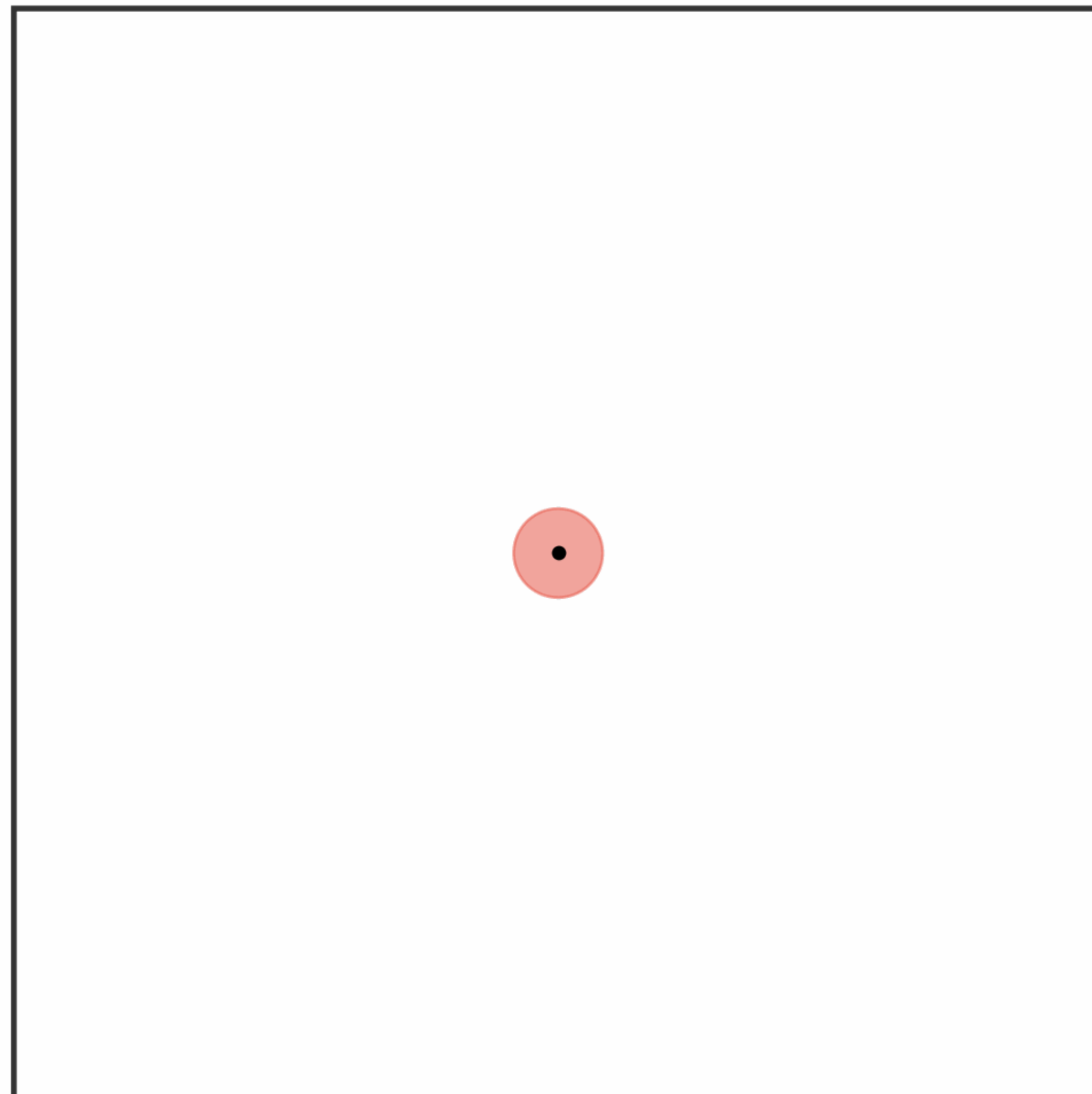
Hyperbolic Geometry

Hyperbolic space $\mathbb{H}^d$

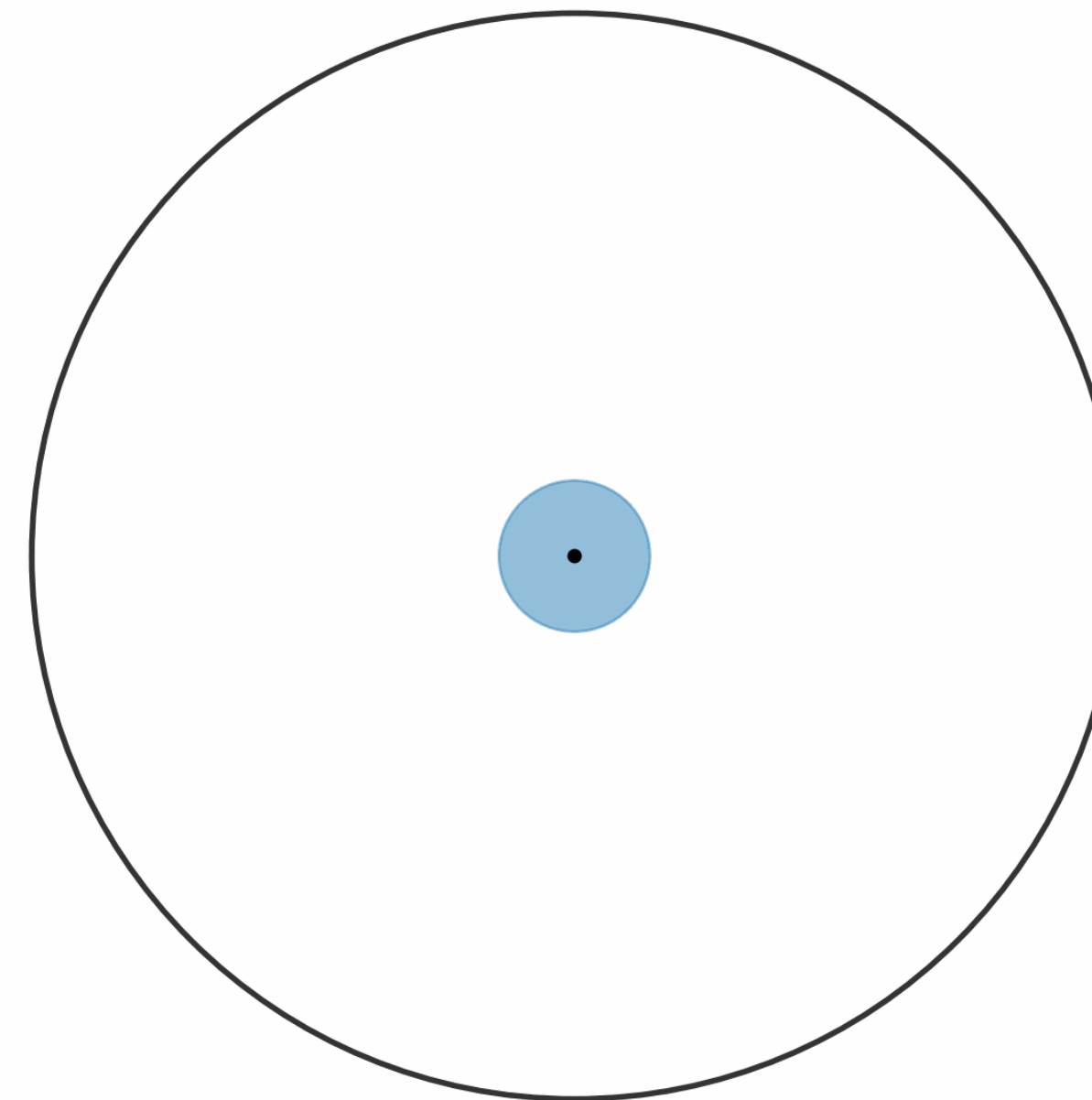
simply connected, negatively curved d -dimensional Riemannian manifold

Hyperbolic VS Euclidean: **volume growth, tree packing**

Euclidean $\mathbb{E}^2$
Polynomial Space ($V \sim r^2$)



Hyperbolic $\mathbb{D}^2$
Exponential Space ($V \sim e^r$)



Hyperbolic Geometry: Motivation

Beyond the geometric interest, hyperbolic geometry gains attention in ML

- Many real-world data has **hierarchical** structures
- Hyperbolic embeddings represents the data better

[NK'17] shows that 5-dim hyperbolic embeddings have lower distortion than 300-dim Euclidean embeddings for WordNet dataset

[He et al. '25] shows that token frequencies in LLMs strictly follow power-law

- Hyperbolic geometry is useful for ML architecture

GNNs, LLMs, etc.

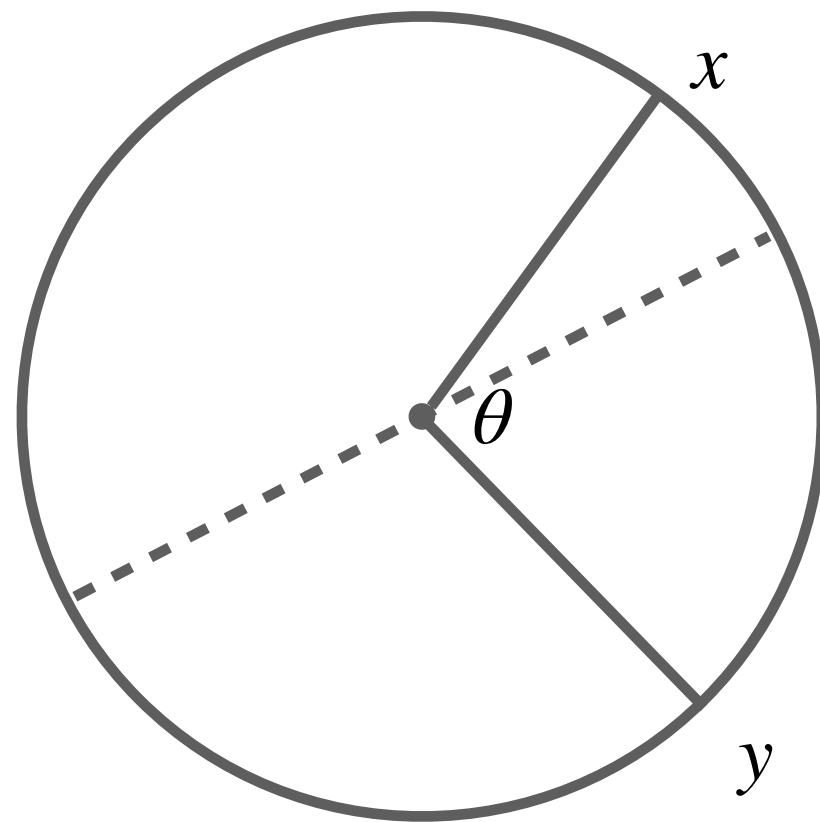


Hyperbolic LSH

Negative signals

- A Riemannian manifold is not a vector space No inner product, [Shrivastava-Li] fails
- No easy regular tessellations exist for high dim No 'grid', [Andoni-Indyk] fails
- Exponential volume growth Far points may be hashed into one bucket

These do not rule out one technique: random hyperplane rounding



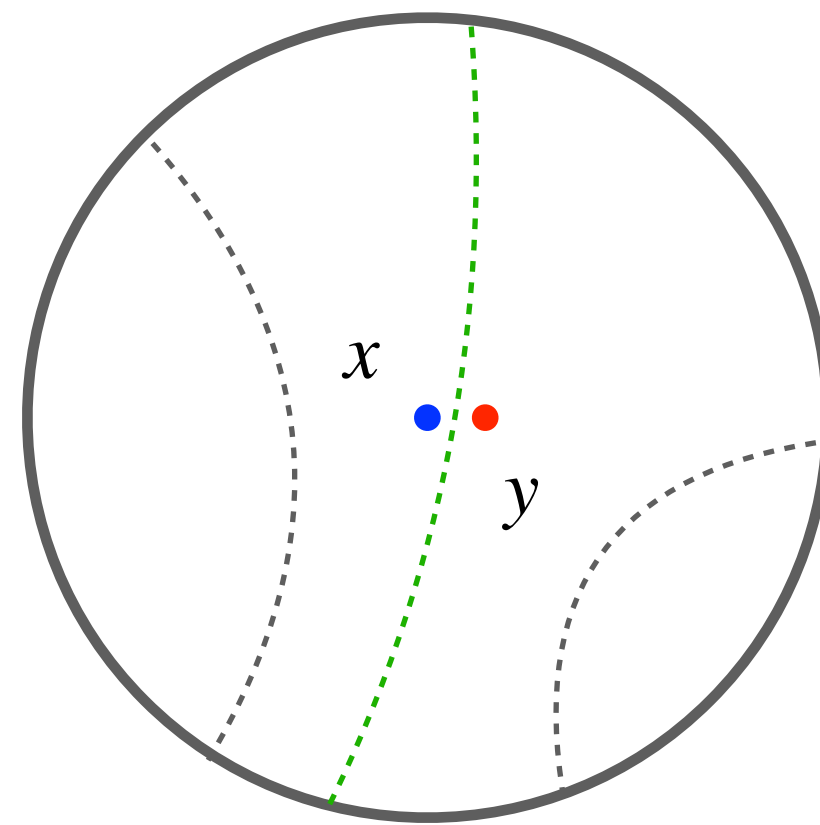
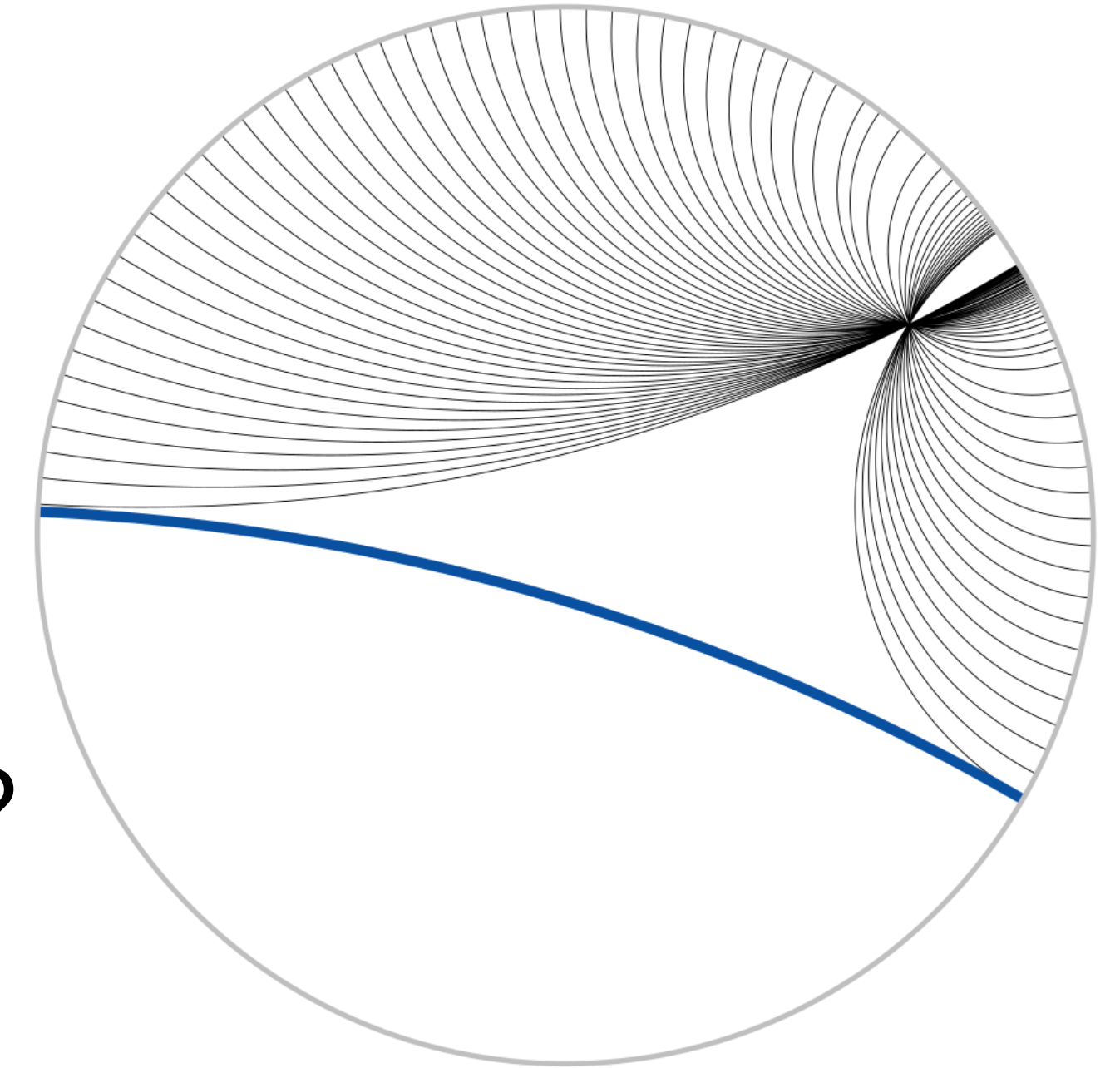
- Uniformly at random sample a straight line l passing through the origin
- If x, y lies on different side of l , $h(x) = 1, h(y) = 0$
- Otherwise, $h(x) = h(y)$

Hyperbolic LSH ($d = 2$)

Poincaré disk model

- Defined on $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$
- Straight lines are **circular arcs orthogonal** to the unit circle

🔍 Can we randomly sample these arcs to separate points?



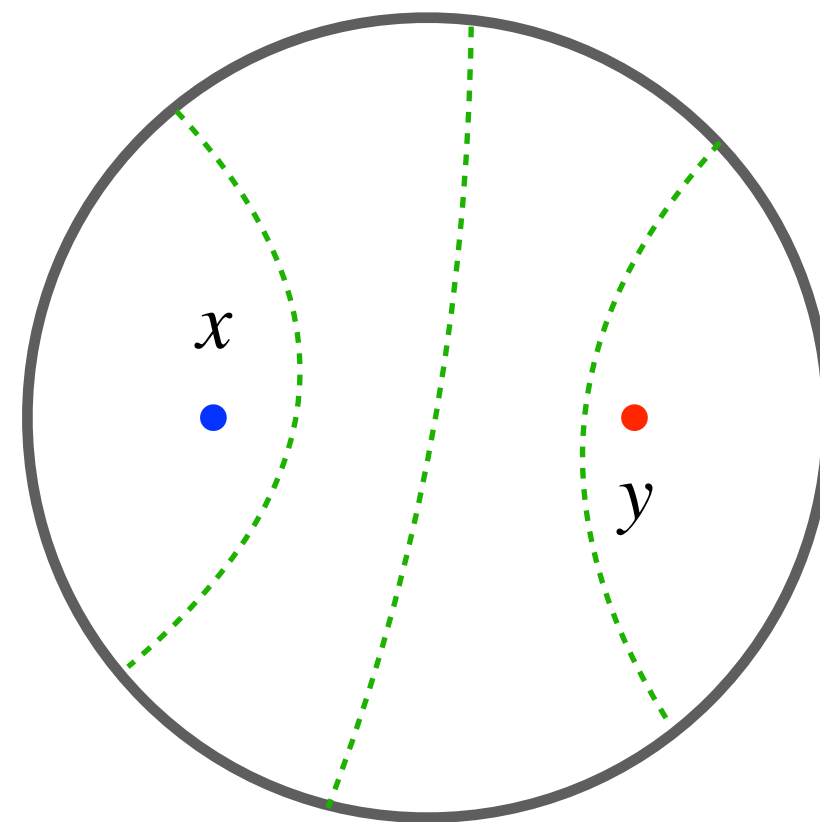
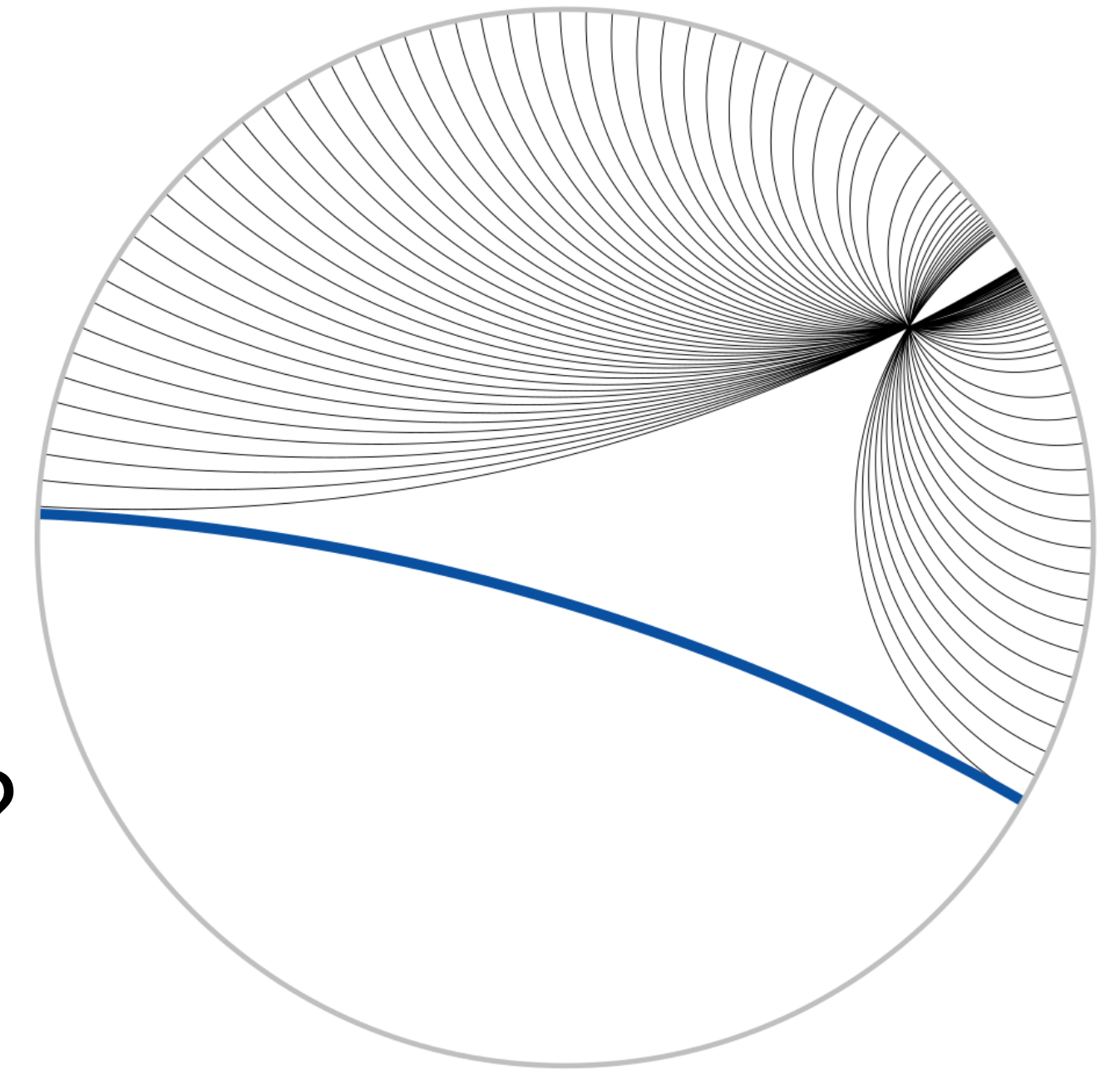
If x, y are close, a random geodesic not likely separate them

Hyperbolic LSH ($d = 2$)

Poincaré disk model

- Defined on $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$
- Straight lines are **circular arcs orthogonal** to the unit circle

🔍 Can we randomly sample these arcs to separate points?



If x, y are close, a random geodesic not likely separate them

If x, y are far, a random geodesic likely separate them

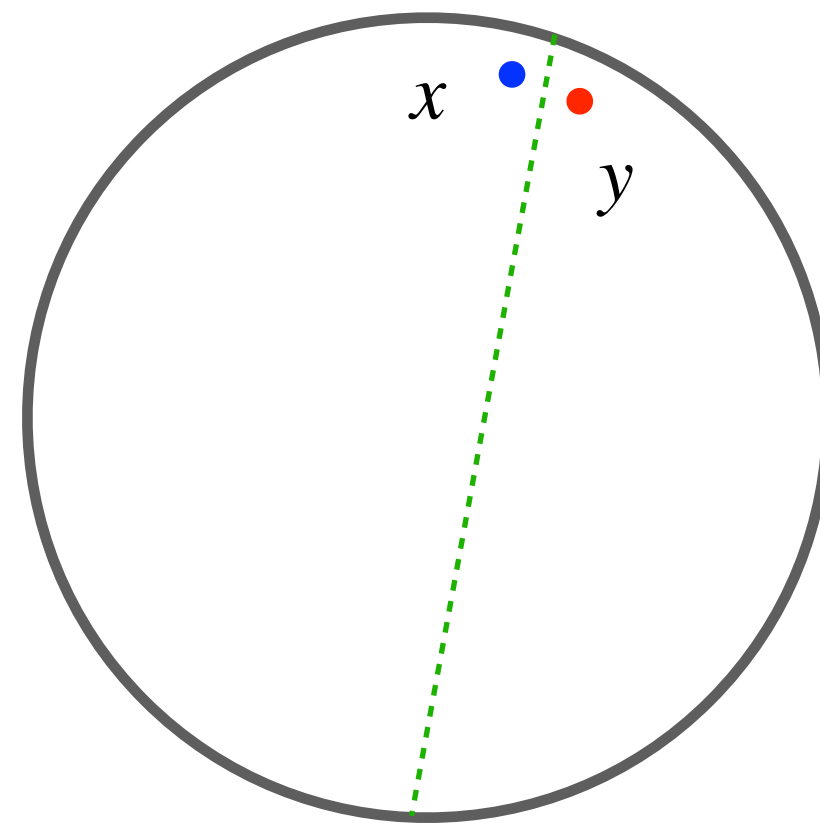
But is this a 'hallucination' of Euclidean?

Hyperbolic LSH ($d = 2$)

Poincaré disk model

- Defined on $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$
- Straight lines are **circular arcs orthogonal** to the unit circle

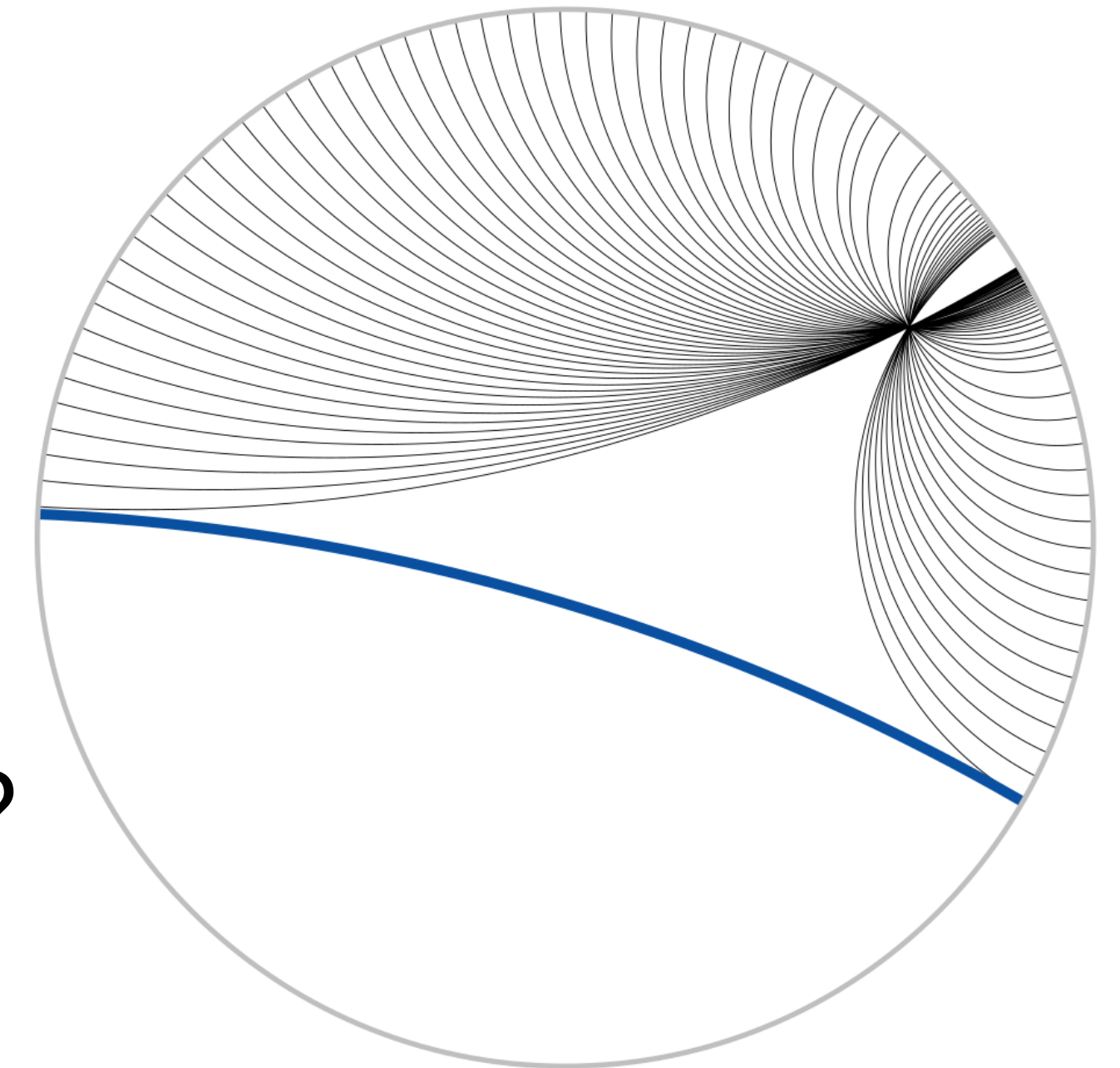
🔍 Can we randomly sample these arcs to separate points?



Here x, y look close but are actually far

The boundary of geodesics curves the same as the space

In this example, many valid separations exist



Hyperbolic LSH ($d = 2$)

LSH function

- Uniformly sample a random geodesic γ
- Hash a point x as $h_\gamma(x) \in \{+1, -1\}$ according to which side of γ it lies on

⚠ Is there something off?

1. To sample, we need a **isometry-invariant measure** on the set of all geodesics

We do have one, **kinematic measure in hyperbolic plane**

2. The space of all geodesics has **infinite total measure**

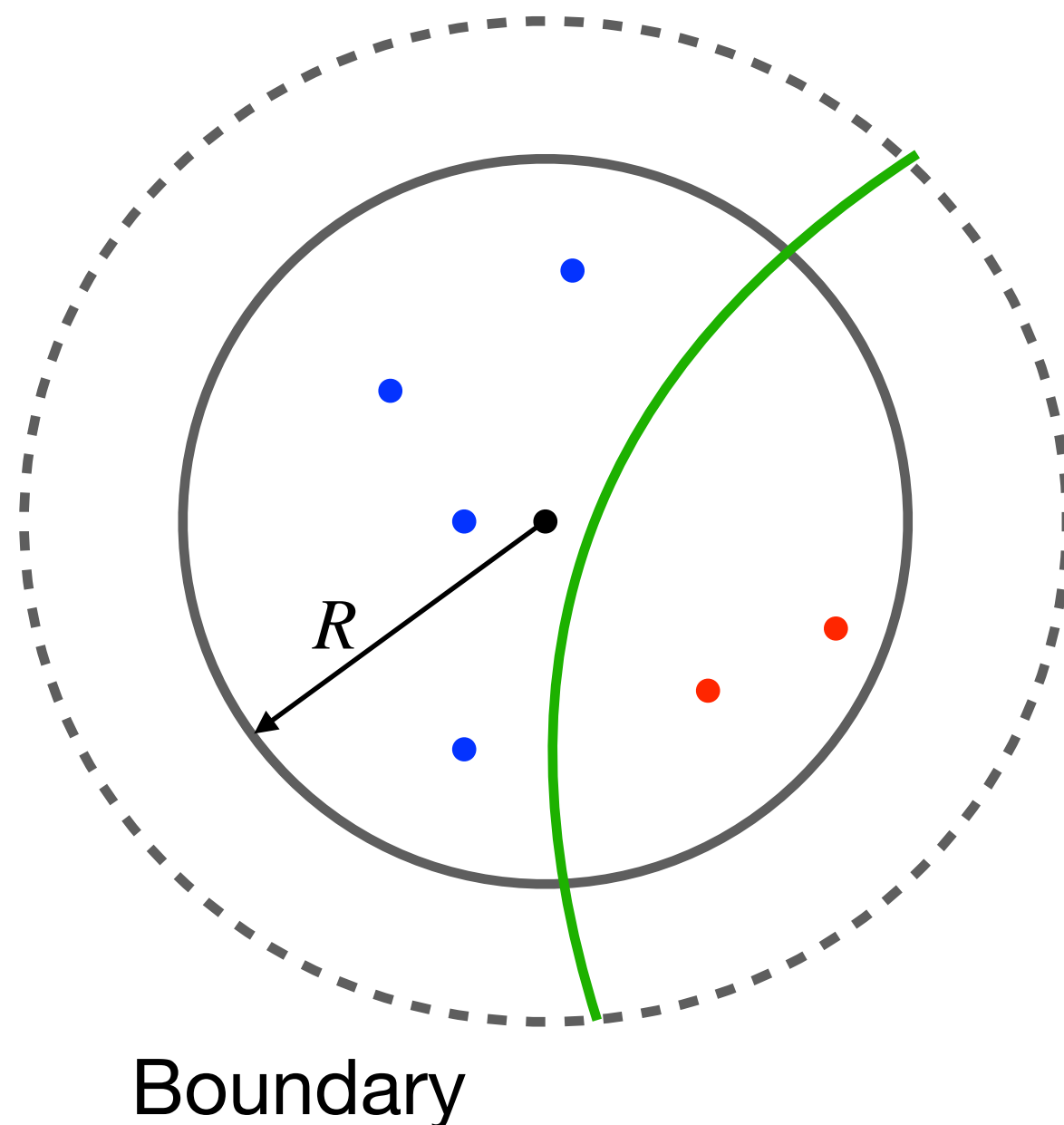
We use a trick, suppose the data is bounded by **radius R**

The analysis also relies on tools from **integral geometry**

Hyperbolic LSH ($d = 2$)

LSH function

- Let R be a radius such that P is contained in the ball $B(0, R)$.
- Let μ_R denote the **normalized** kinematic measure of the set of geodesics intersecting $B(0, R)$
- Uniformly sample a random geodesic $\gamma \sim \mu_R$
- Hash a point x as $h_\gamma(x) \in \{+1, -1\}$ according to which side of γ it lies on



By calculation

$$\mu_R(\gamma) = \frac{1}{2\pi \sinh(R)} d\mu_{\text{kinematic}}$$

$$d\mu(t, \theta) = \cosh(t) dt d\theta$$

Theorem 1. The family H of above function yields $\rho(H) \leq 1/c$

Hyperbolic LSH Analysis ($d = 2$)

Key lemma (Collision probability)

$$\Pr[h(x) = h(y) \mid d_{\mathbb{H}}(x, y) = r] = 1 - \frac{r}{\pi \sinh(R)}$$

Together with the following lemma:

$$\text{For } x \in [0, 1) \text{ and } \ell > 1 \text{ such that } 1 - \ell x > 0, \frac{\log(1 - x)}{\log(1 - \ell x)} \leq \frac{1}{\ell}$$

✓ We are done

$$p_1 = 1 - \frac{r}{\pi \sinh(R)}$$

$$p_2 = 1 - \frac{cr}{\pi \sinh(R)}$$

$$\rho = \frac{\log p_1}{\log p_2}$$

Proof of Key Lemma

Open up the collision probability

$$\Pr[h(x) \neq h(y) \mid d_{\mathbb{H}}(x, y) = r] = \frac{\mu(X)}{\mu(Y)}$$

where X is the set of all geodesics separating x, y

Y is the set of all geodesics intersecting $B(0, R)$

Same standard calculation, $\mu(Y) = \int_0^{2\pi} \int_0^R \cosh(t) dt d\theta = 2\pi \sinh R$

Next we show $\mu(X) = 2r$, where $d_{\mathbb{H}}(x, y) = r$

The total measure of geodesics separating 2 points is **linear** in their distance

Proof of Key Lemma

Analyzing a geodesic separates x, y

- OAy is a right triangle, by [hyperbolic cosine law](#)

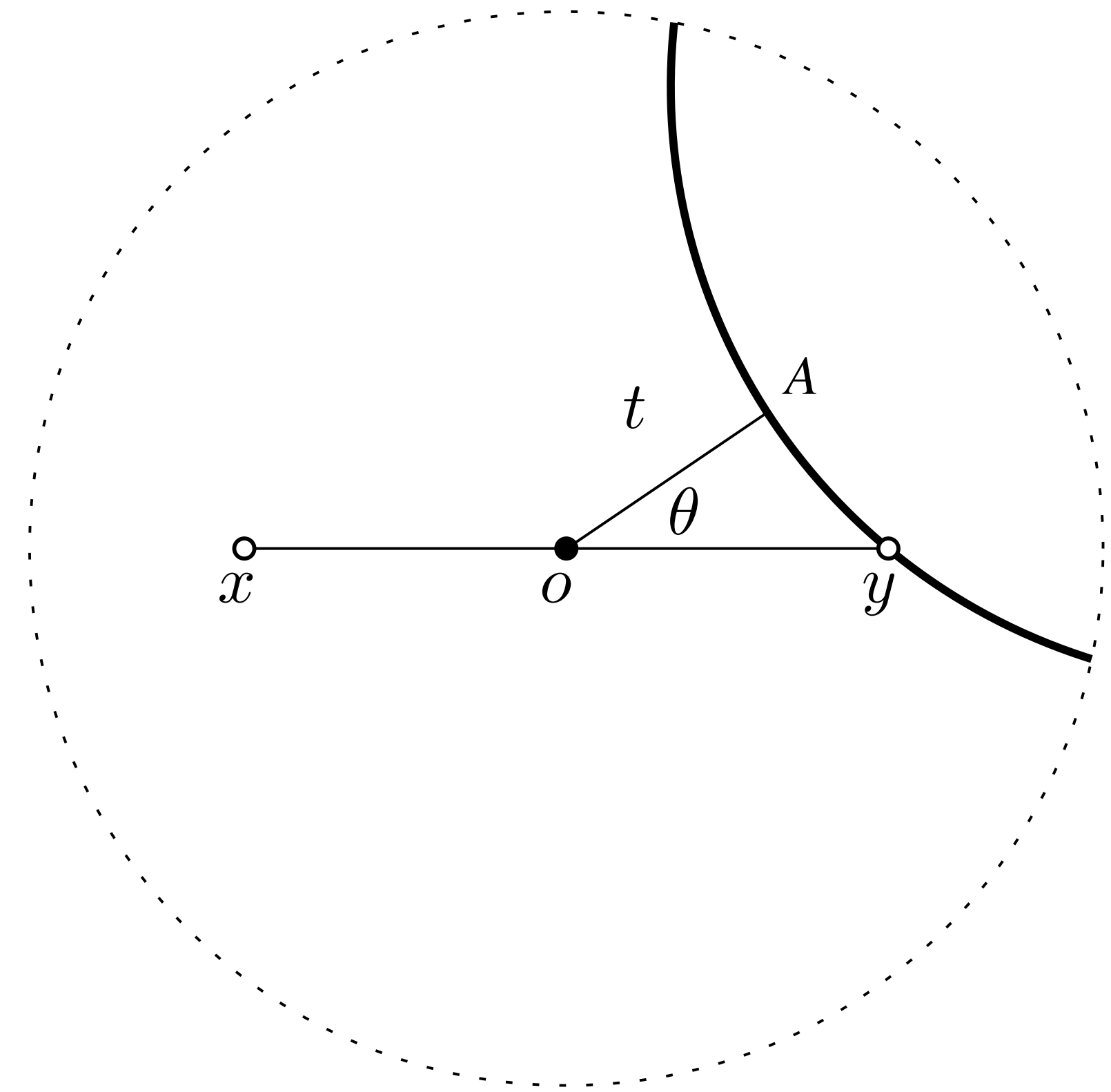
$$\tanh(t) = \tanh\left(\frac{r}{2}\right)\cos(\theta)$$

- X in the [polar coordinate](#) is

$$X = \left\{ (t, \theta) \mid \theta \in [-\pi, \pi], 0 \leq t \leq \operatorname{arctanh}\left(\tanh\left(\frac{r}{2}\right)\cos \theta\right) \right\}$$

- Deriving the [Crofton Formula](#)

$$\mu(X) = 2 \int_{-\pi/2}^{\pi/2} \int_0^{\operatorname{arctanh}\left(\tanh\left(\frac{r}{2}\right)\cos \theta\right)} \cosh(t) dt d\theta = 2r$$



Hyperbolic LSH (high-dim)

If you wonder

Why not use **high-dim kinematic measure?** (Sample random hyperplanes)

$$d\mu(t, \theta) = \cosh^{d-1}(t) dt d\theta$$

Maybe, but the algebra is very complicated

We look into **dimension reduction** in hyperbolic space

Can we project $\mathbb{H}^d$ to $\mathbb{H}^2$ with low (constant) distortion?

If yes, we combine dimension reduction with the 2d LSH

Hyperbolic Dimension Reduction

Theorem 2 (Informal)

Let $k \in \mathbb{N}, k \geq 1$, if there exists an (α, β) -approximate dimension reduction for an n -dimension Euclidean point set to k -dimension Euclidean space, then there exists an (α, β) -approximate dimension reduction for a $(n + 1)$ -dimension hyperbolic point set to $(k + 1)$ -dimension hyperbolic space.

(α, β) -approximate

There exists a mapping f that projects the input X into $(\hat{X}, \hat{d})$, such that

$$\alpha \cdot d(p, q) \leq \hat{d}(f(p), f(q)) \leq \beta \cdot d(p, q)$$

 Hyperbolic space also adopts a **JL lemma**

[Datar-Immorica-Indyk-Mirroknii'04]

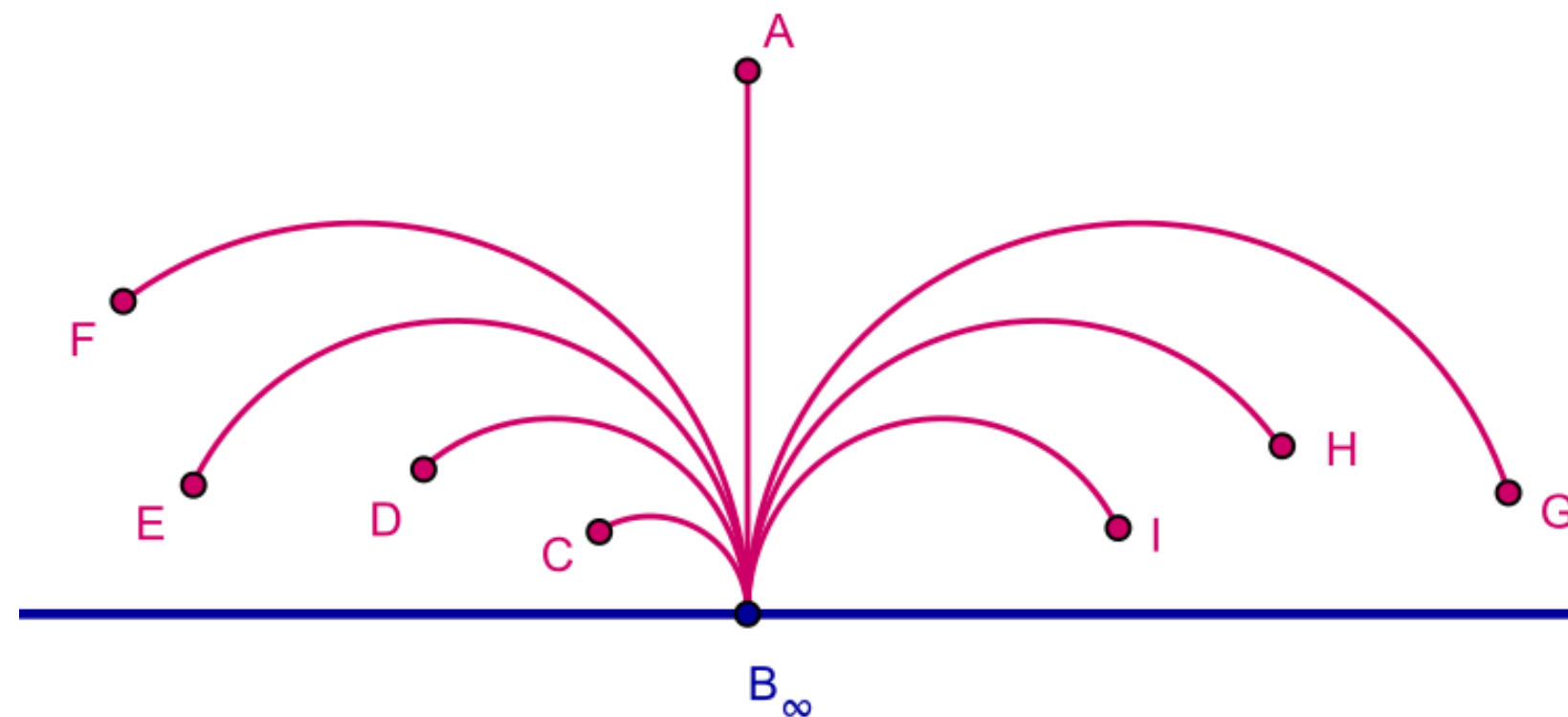
Let $k = 1$, if we know how to project Euclidean n -space to **a line**, we are good

Poincaré Half-space Model

Each point in $\mathbb{H}^d$ is (z, x) where $z \in \mathbb{R}^+$ and $x \in \mathbb{R}^{d-1}$

- x is the $(d - 1)$ -dim Euclidean “projection”
- z is the “depth” coordinate, infinity at $z \rightarrow 0$, logarithmic space expansion
- Distance function

$$d_{\mathbb{H}}(p_1, p_2) = \operatorname{arccosh} \left(1 + \frac{\|x_1 - x_2\|_2^2 + (z_1 - z_2)^2}{2z_1z_2} \right)$$



Hyperbolic Dimension Reduction

Main idea: z is the 'lazy' dimension, just **keep z there**; Apply **Euclidean dimension reduction to x** .

The idea is given by [Benjamini-Makarychev'07]

We only refined the analysis

💡 The result may have broader interest

With one extra dimension cost, we gain same dimension reduction in hyperbolic space as Euclidean space

Hyperbolic LSH (high-dim)

LSH function

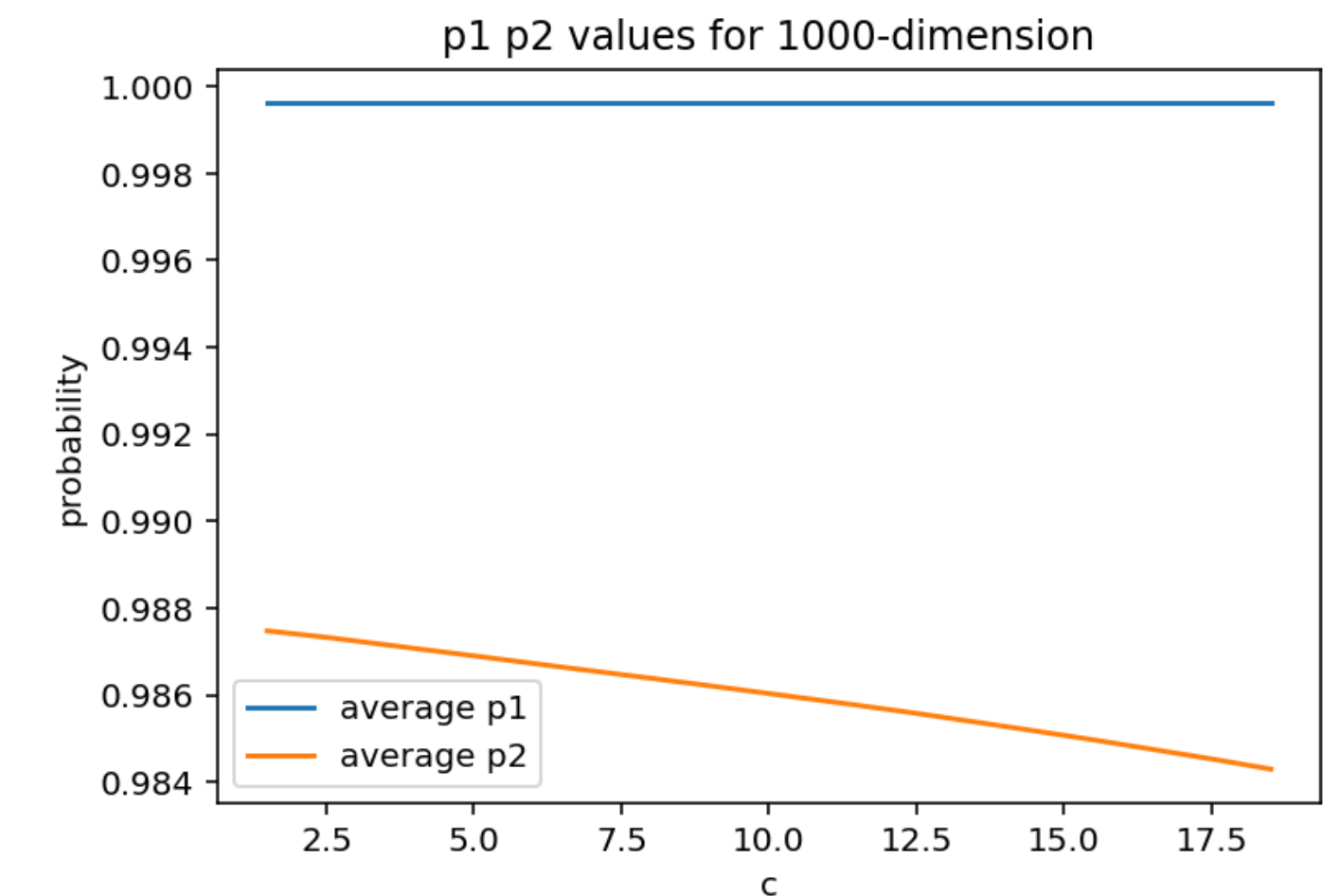
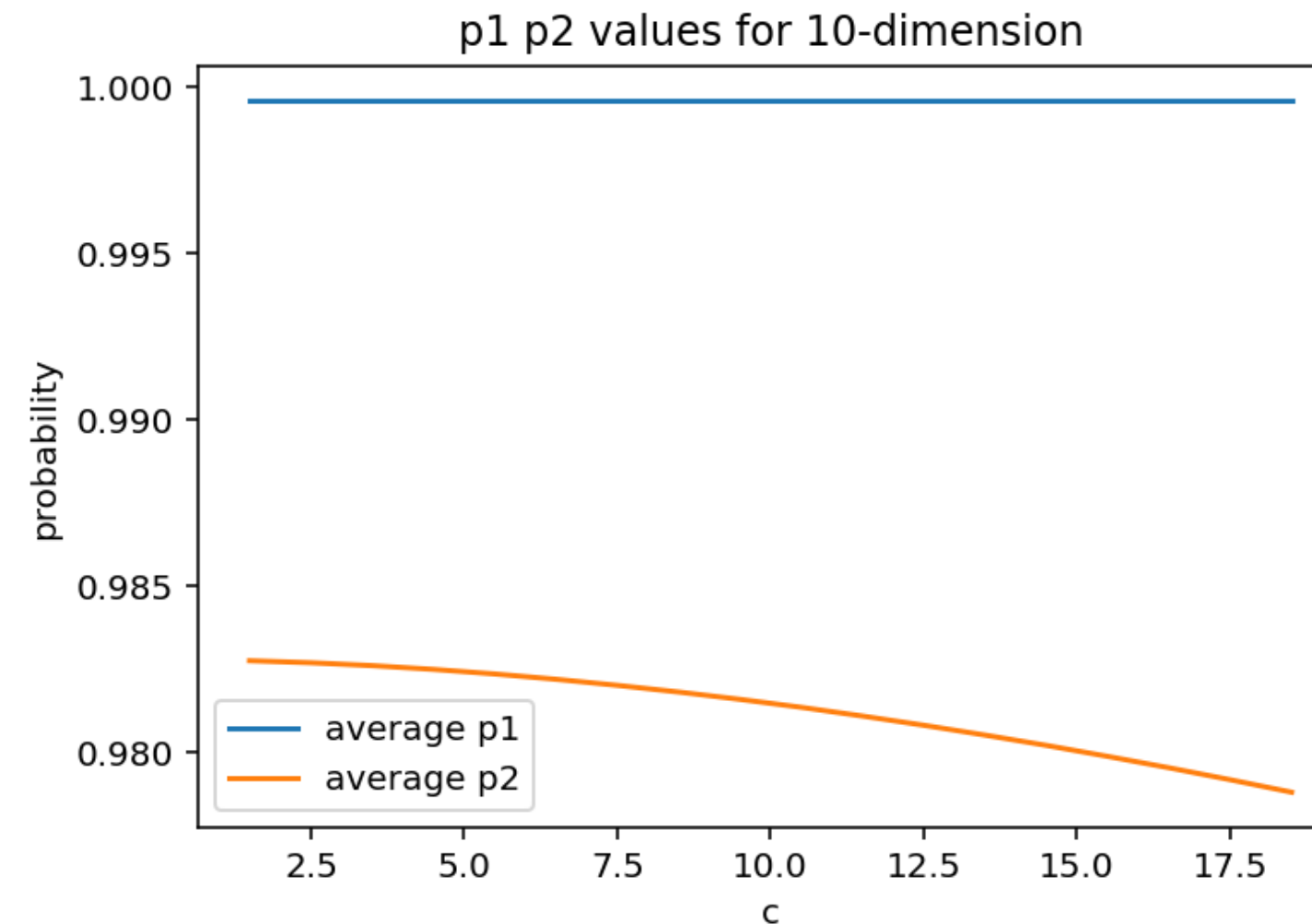
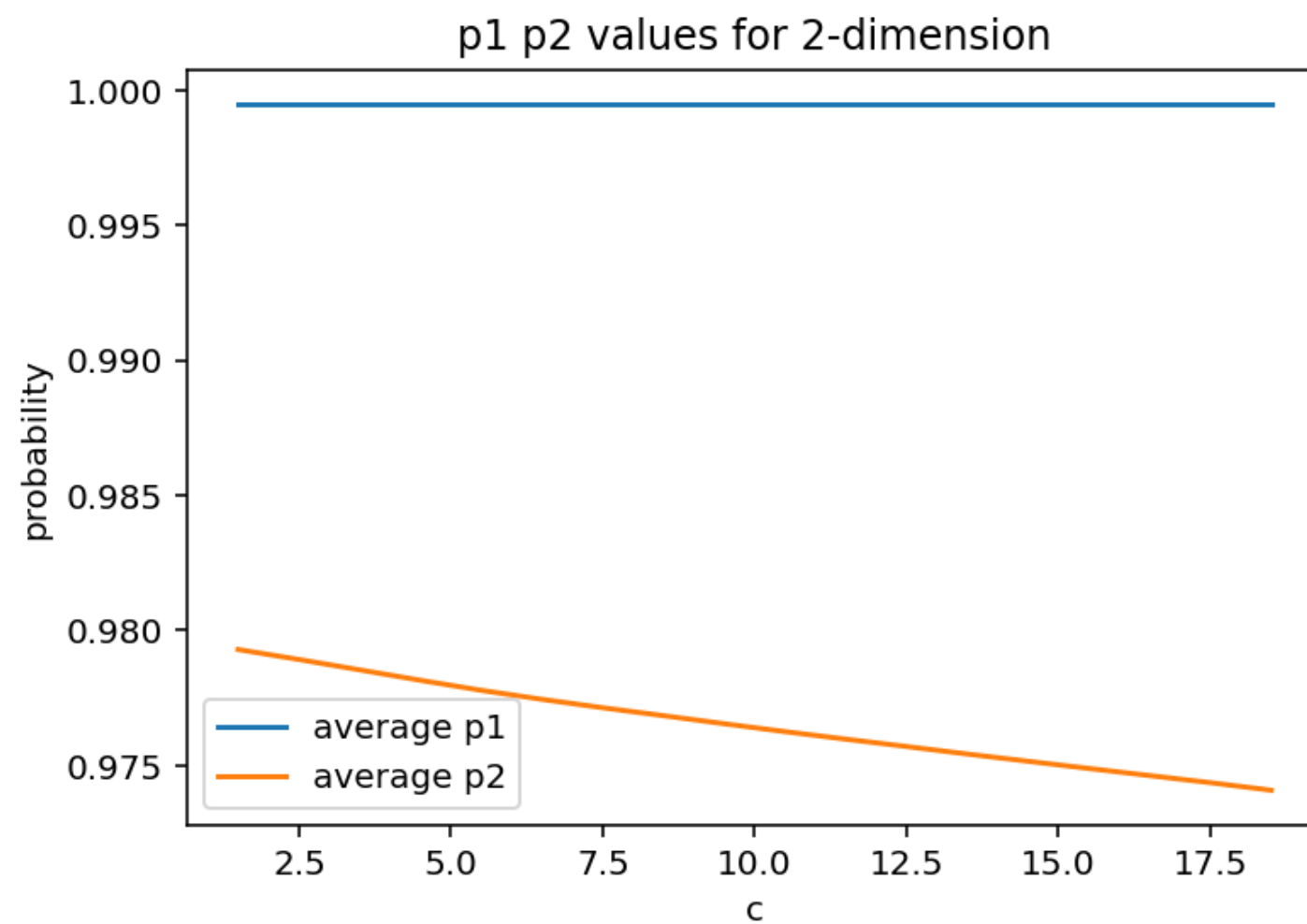
- Dimension reduction, $\mathbb{H}^d \rightarrow \mathbb{H}^2$
Keep z , for the $(d - 1)$ -dim Euclidean ‘subspace’, project it onto a line
- Apply the random geodesic rounding scheme in $\mathbb{H}^2$

Theorem 3. The family H of above function yields $\rho(H) \leq 1.59/c$

Hyperbolic LSH Experiments

Setup

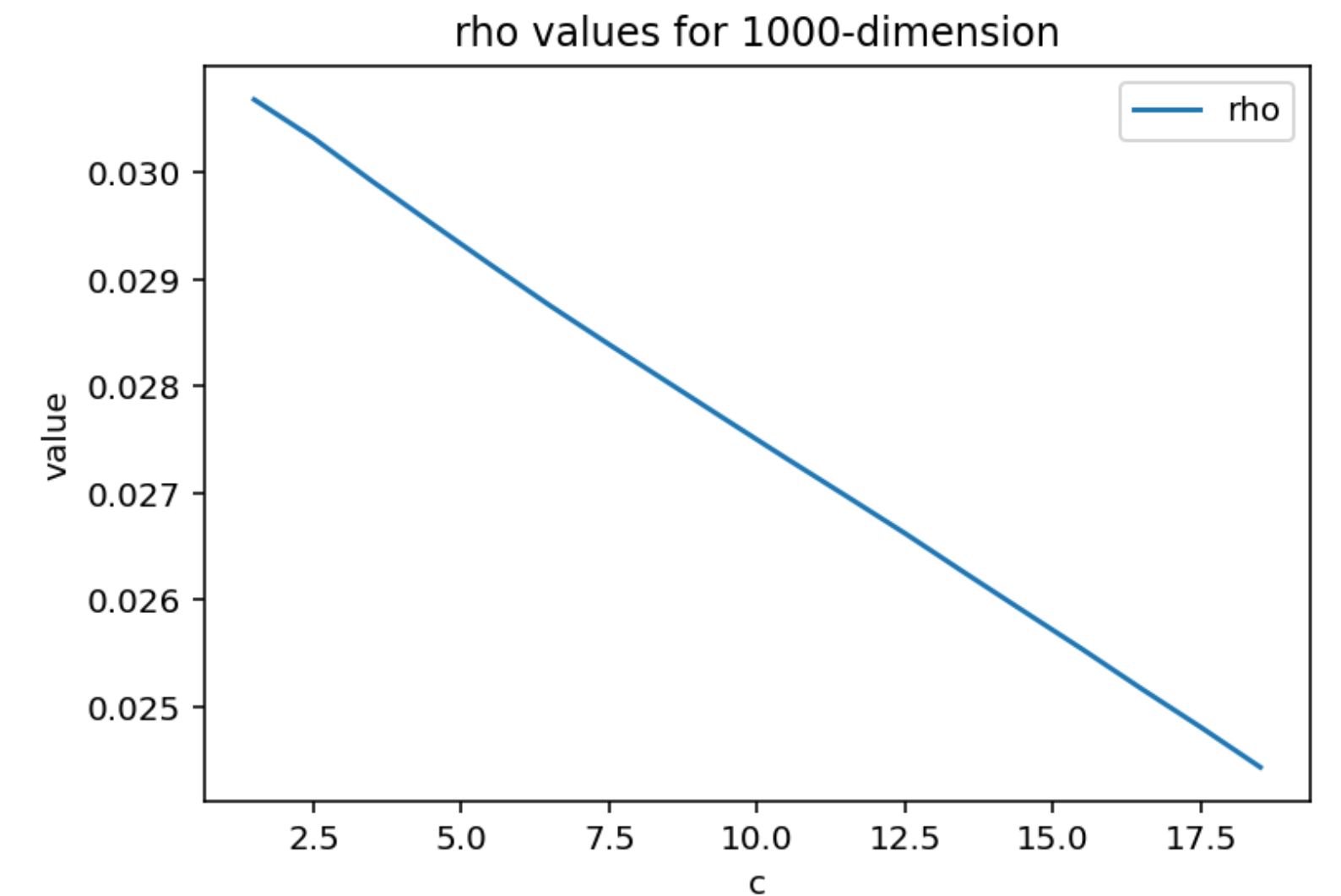
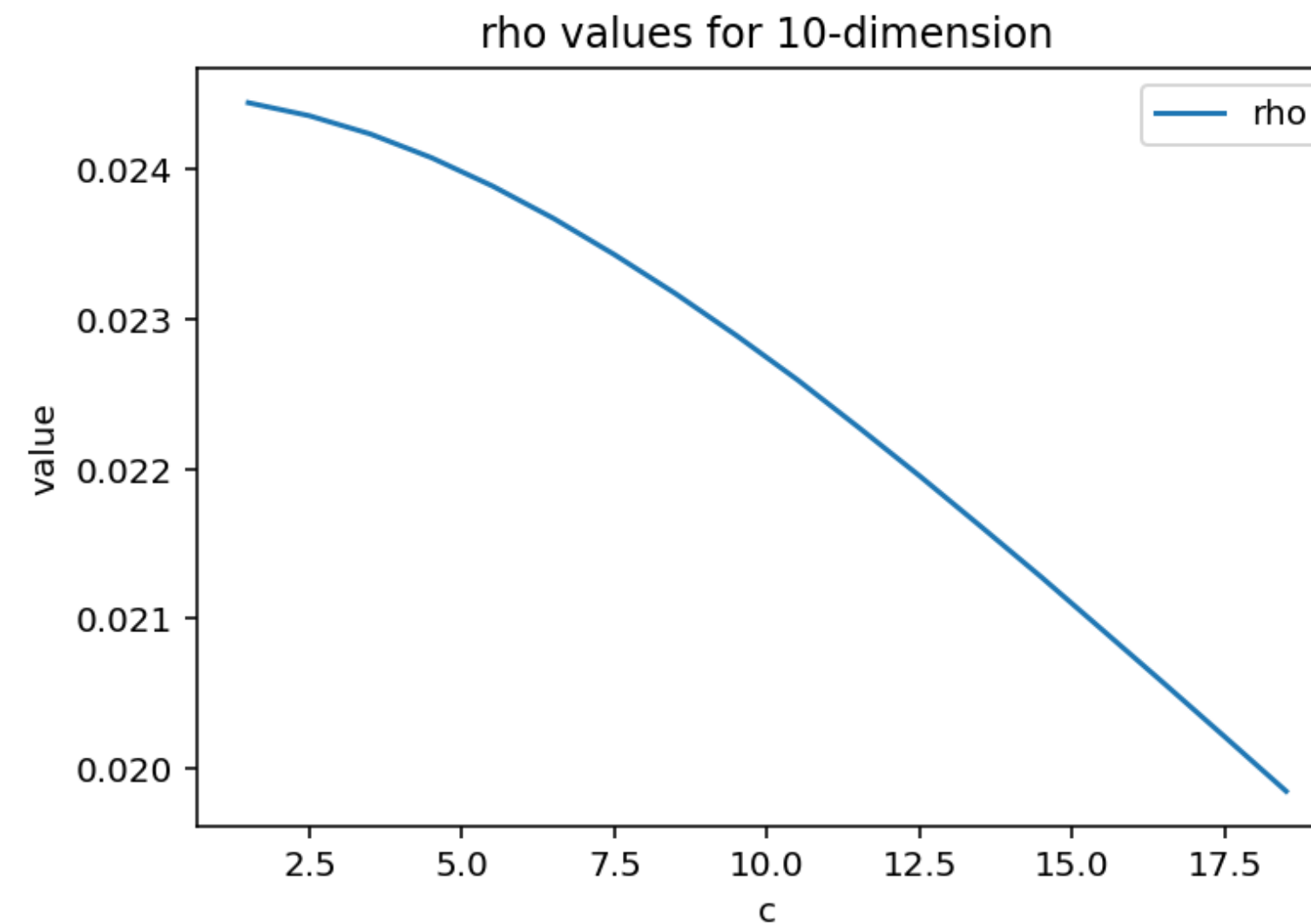
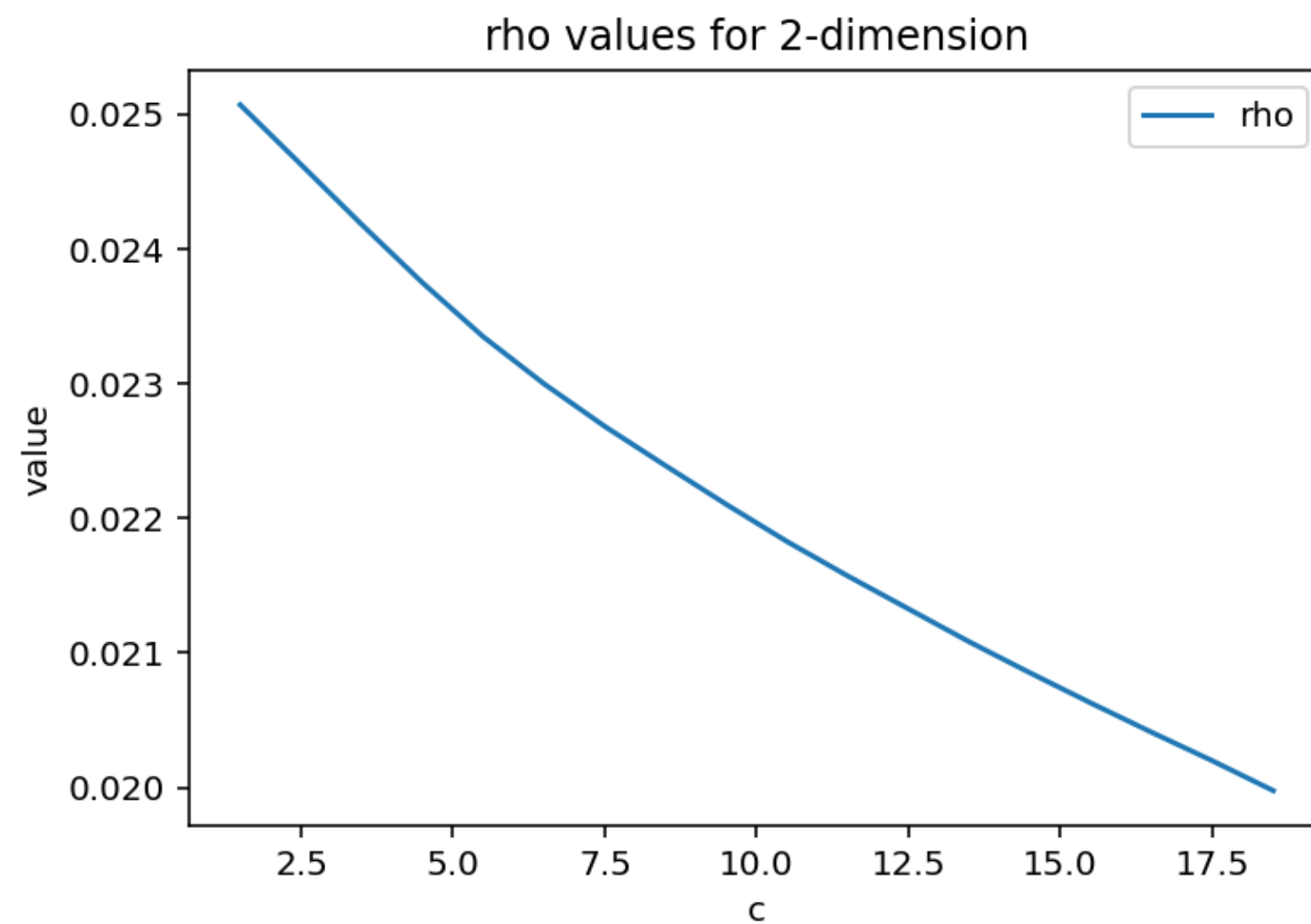
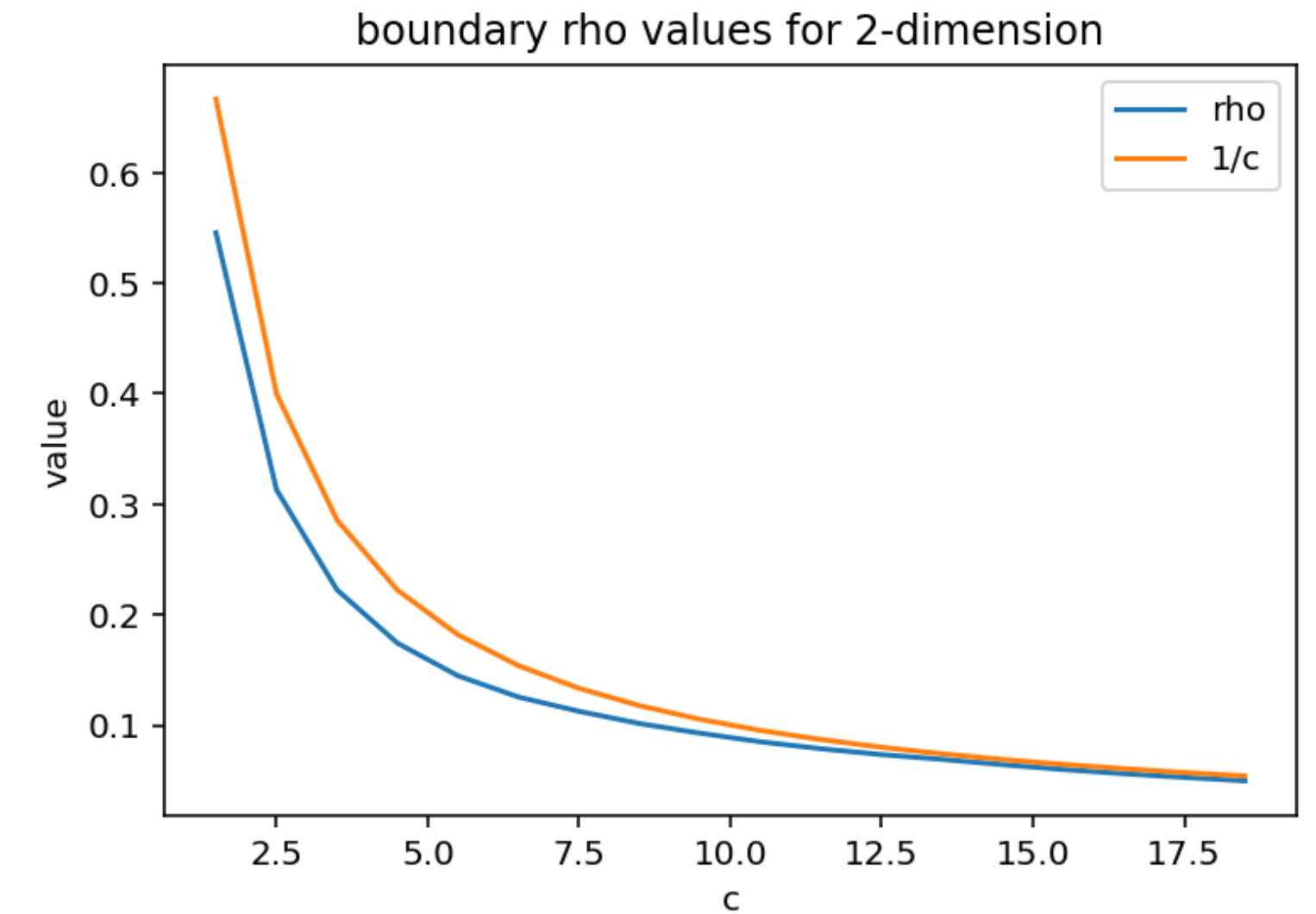
- Dimension $d = \{2, 10, 1000\}$
- Radius $R = \log 199$, $r = 0.2$, $c = 1.5 + k$
- 1000 points, one hash with 1000 repetitions



Hyperbolic LSH Experiments

Setup

- Dimension $d = \{2, 10, 1000\}$
- Radius $R = \log 199$, $r = 0.2$, $c = 1.5 + k$
- 1000 points, one hash with 1000 repetitions



Summary

- We give data-independent LSH for hyperbolic space
- For plane, we borrow the random hyperplane rounding scheme
- For high-dim, we combine dimension reduction with an existing Euclidean LSH
- A JL-type lemma exists for Hyperbolic space

Future Directions

- Close the gap, both plane and high-dim
- If we allow data-dependent, any improvement?
- We move into a new geometry (quite different from Euclidean)
 - What can be done, what cannot, or what is the cost?
 - **YES:** JL, LSH, Quadtree [Kisfaludi-Bak and Wordragen'24], ... ?
 - **NO:** Locality sensitive ordering, ... ?
- For ANNS, navigable graph recently gains more attention
 - Build a graph native to hyperbolic space, to allow for efficient greedy routing

Thank you !