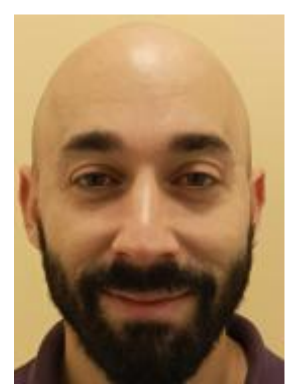
Bridging Arbitrary and Tree Metrics via Differentiable Gromov Hyperbolicity


Pierre Houedry


Nicolas Courty

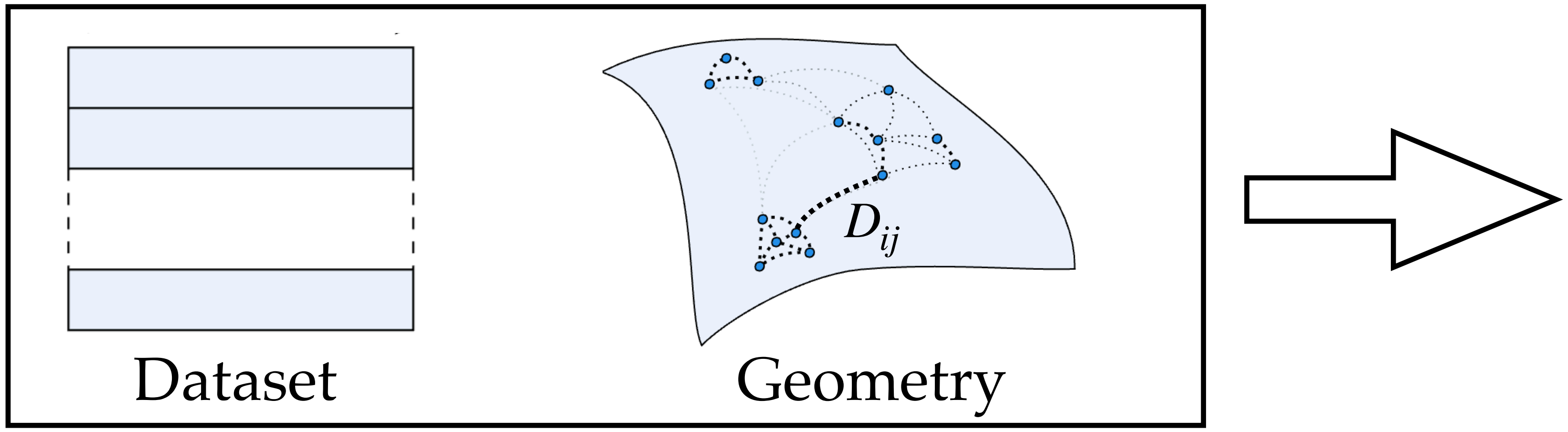

Florestan Martin-Baillon


Laetitia Chapel


Titouan Vayer

Motivations

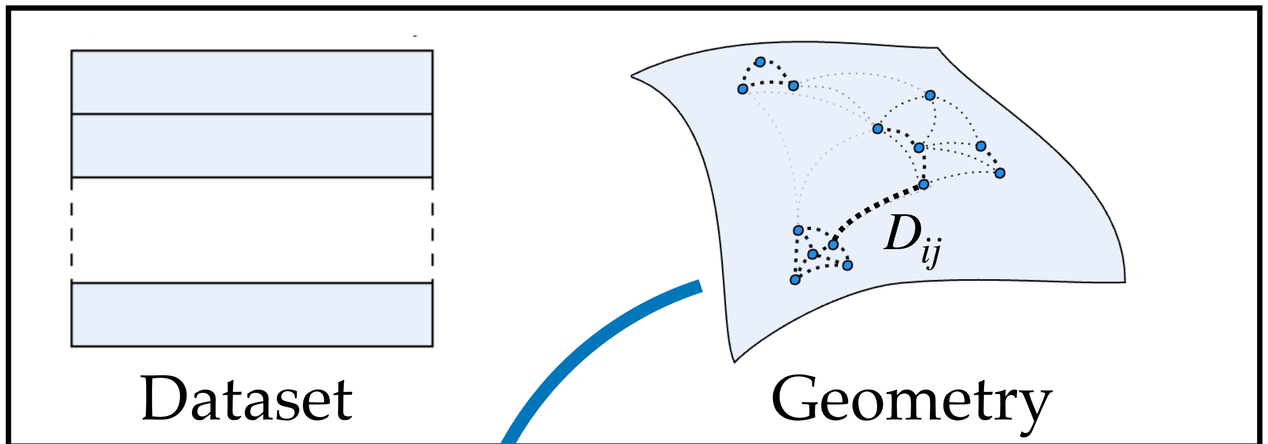
✦ Representation learning



« Interesting » embedding space

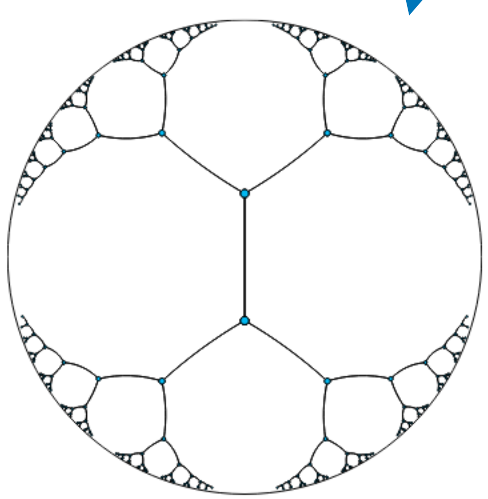
Motivations

✦ Representation learning



« Interesting » embedding space

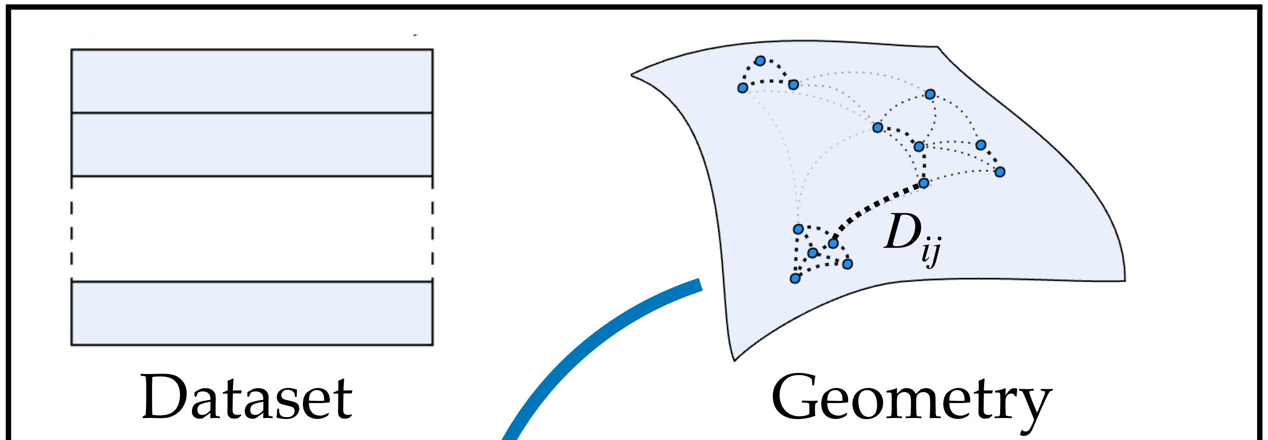
✦ Hierarchical representations



Trees embedding

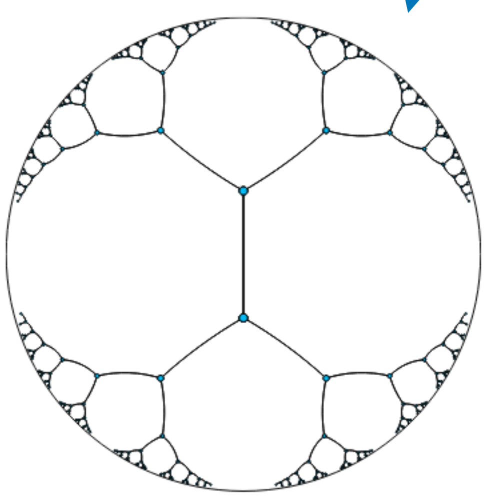
Motivations

✦ Representation learning

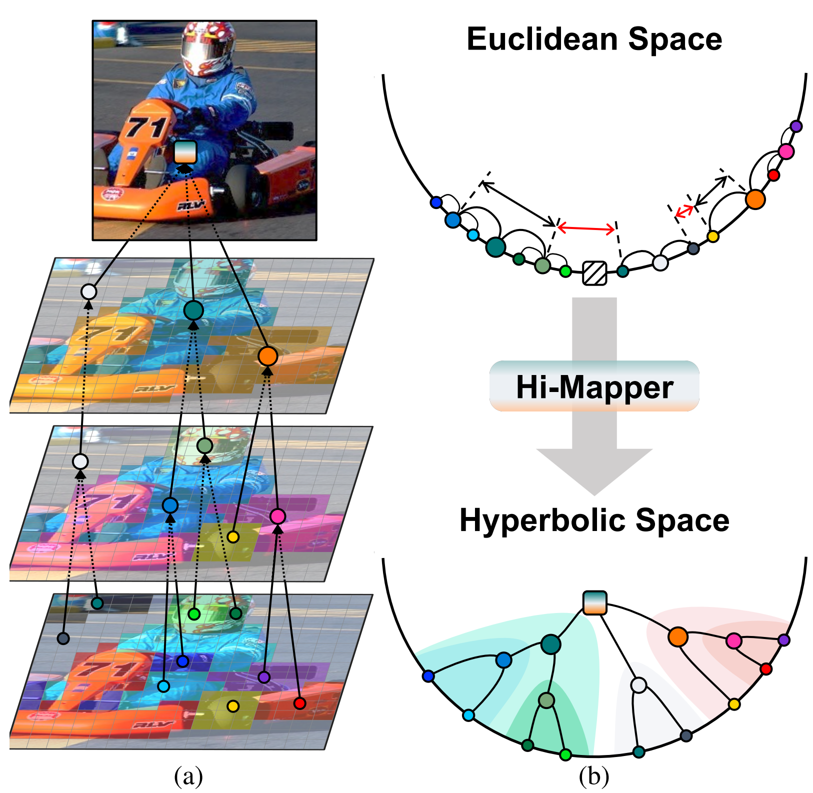


« Interesting » embedding space

✦ Hierarchical representations

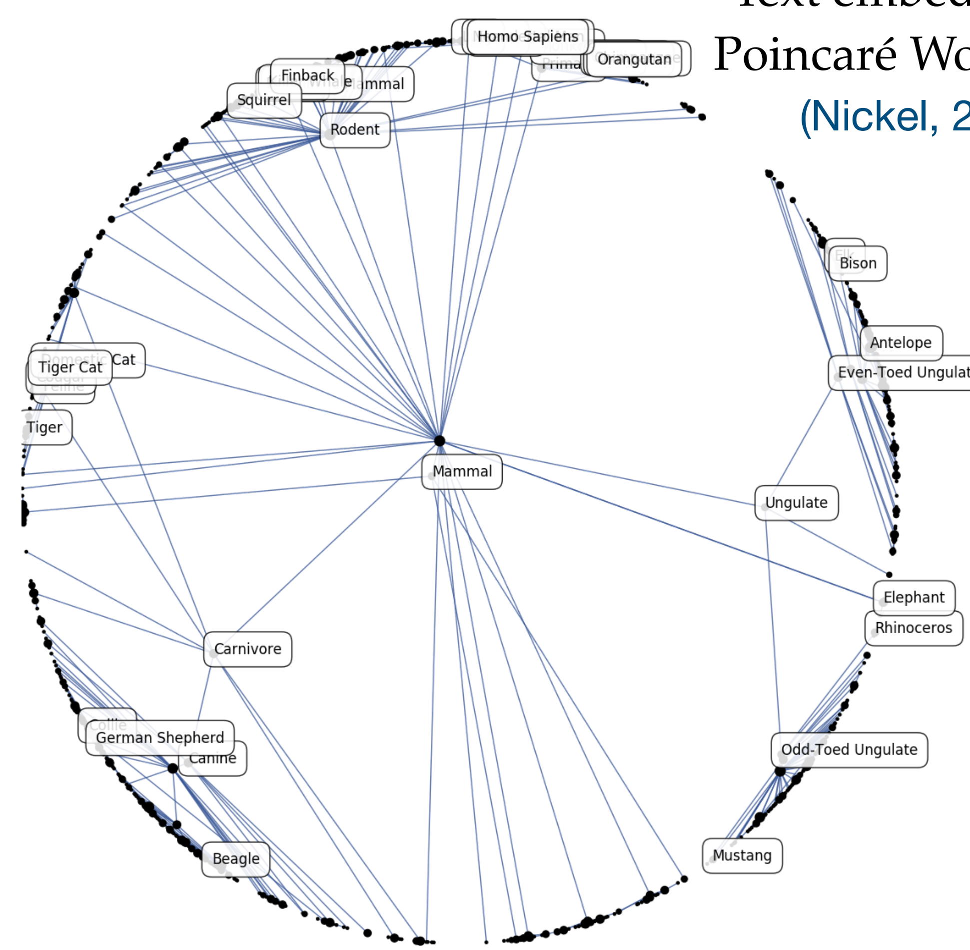


Trees embedding

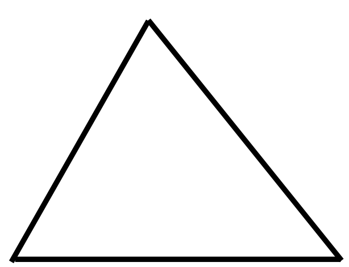


Computer vision (Kwon, 2024)

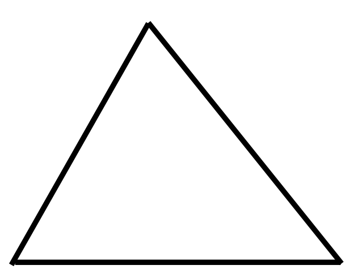
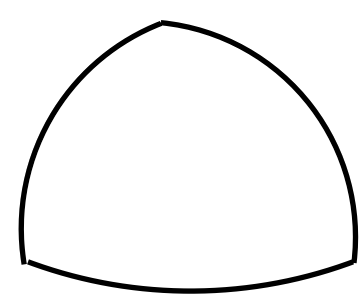
Text embedding: Poincaré WordNet (Nickel, 2017)



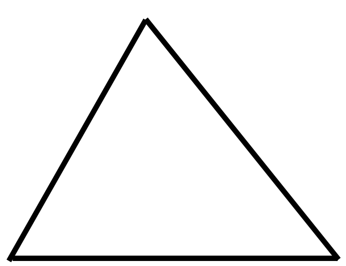
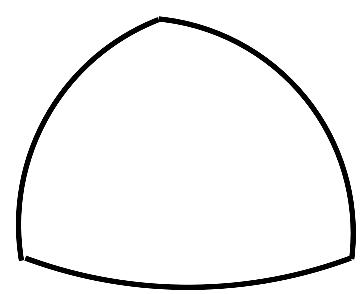
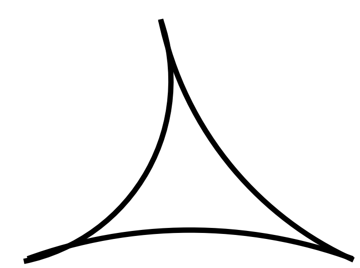
Geometries

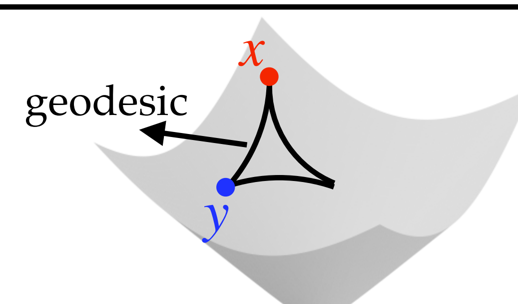
	Euclidean
Curvature	0
Parallel lines	1
Triangles are	normal
Shape	

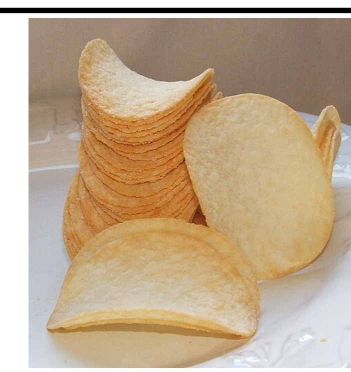
Geometries

	Euclidean	Spherical
Curvature	0	> 0
Parallel lines	1	0
Triangles are	normal	thick
Shape		

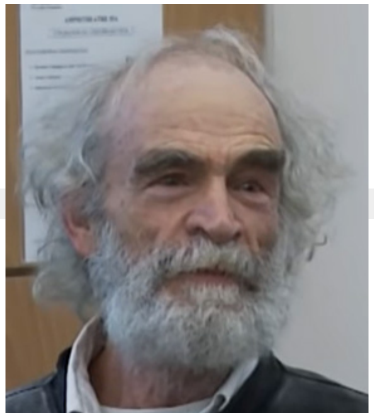
Geometries

	Euclidean	Spherical	Hyperbolic
Curvature	0	> 0	< 0
Parallel lines	1	0	∞
Triangles are	normal	thick	thin
Shape			



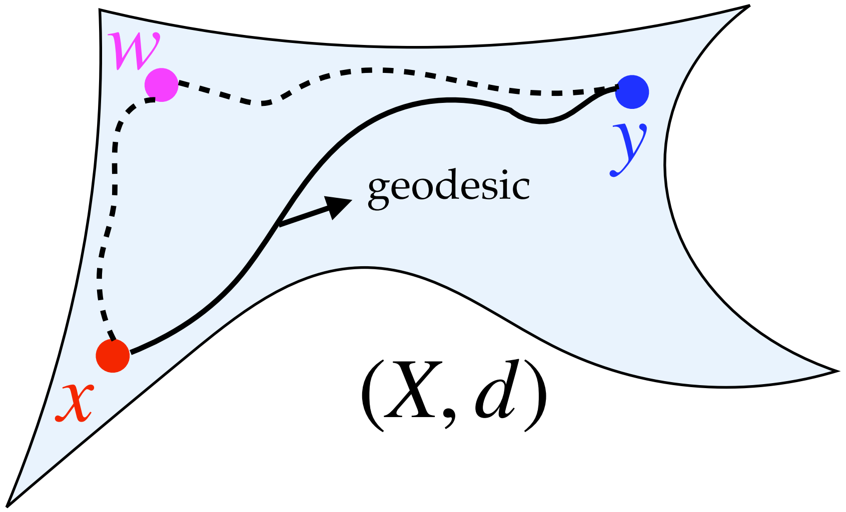


Gromov hyperbolicity



✦ Gromov product

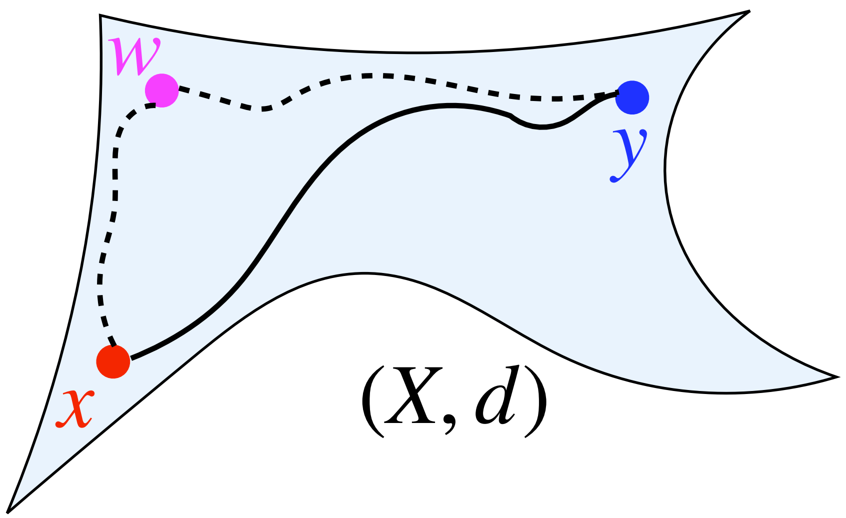
$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



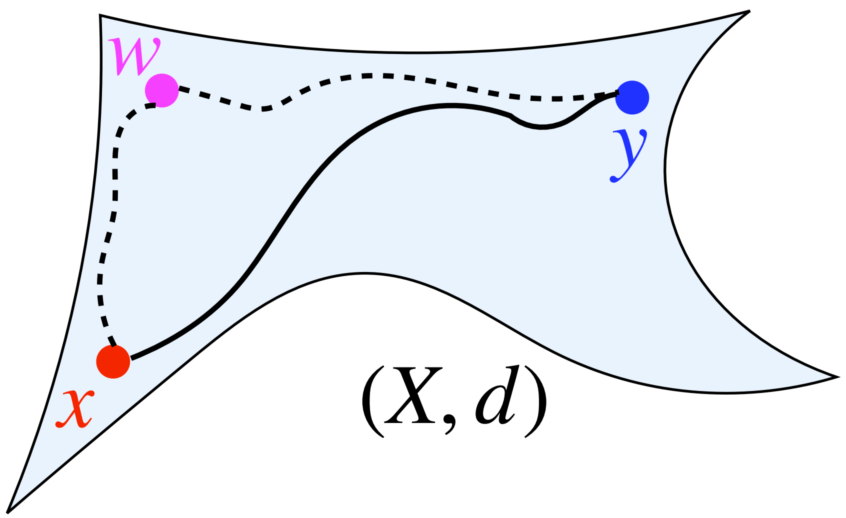
$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

car $(x\,|\,y)_w \le \frac{1}{2}\big(d(x,y) + d(y,w) + d(y,w) - d(x,y)\big)$

Gromov hyperbolicity

✦ Gromov product

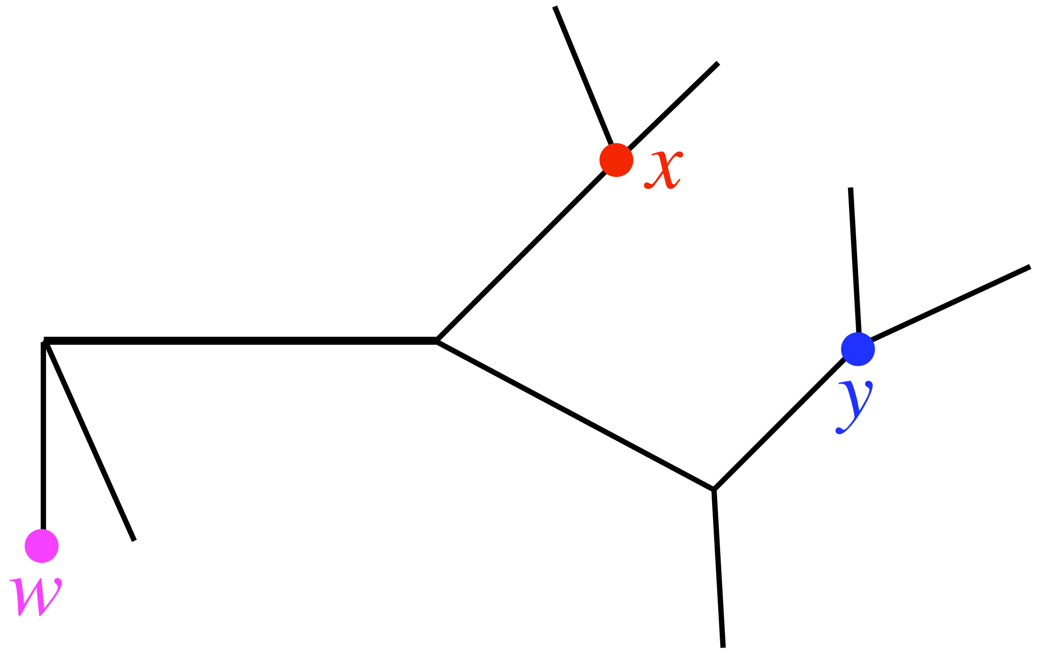
$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$

✦ Interpretation

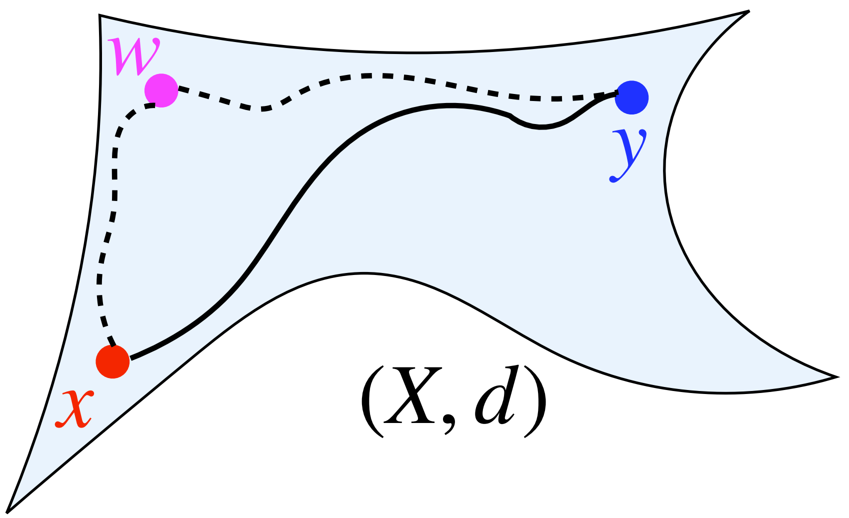
$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging



Gromov hyperbolicity

✦ Gromov product

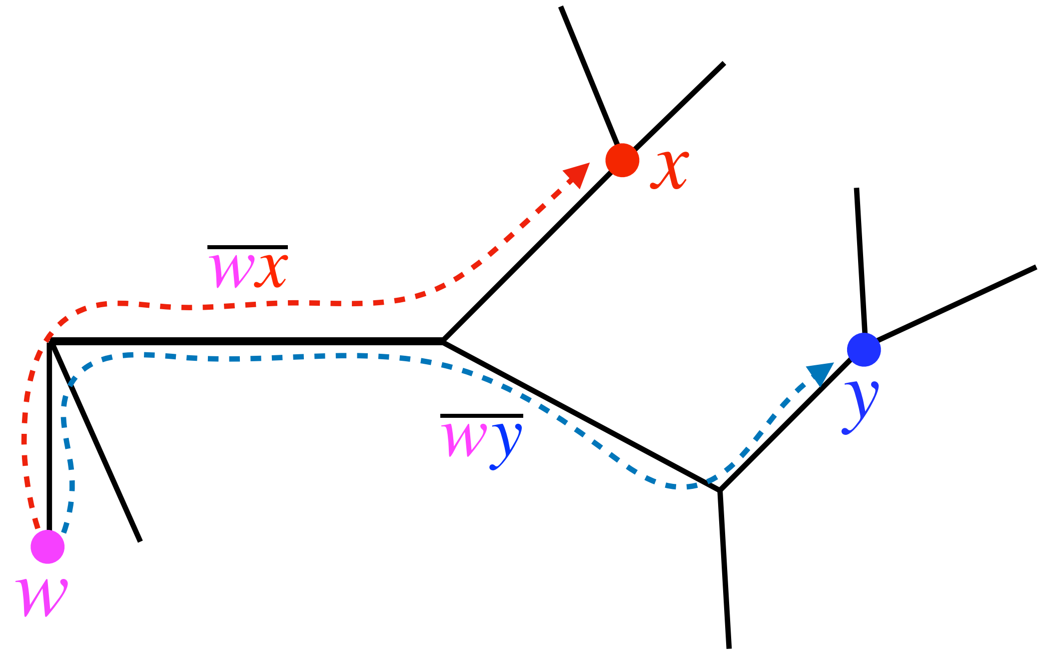
$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Interpretation

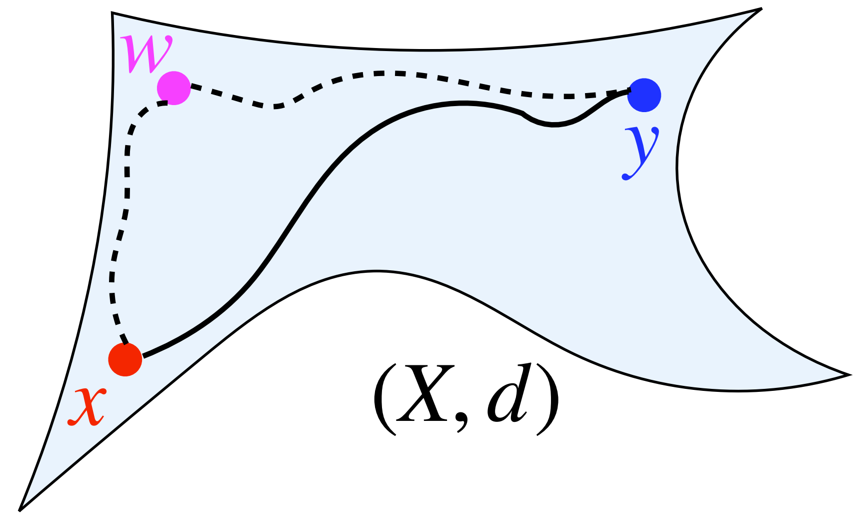
$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging



Gromov hyperbolicity

✦ Gromov product

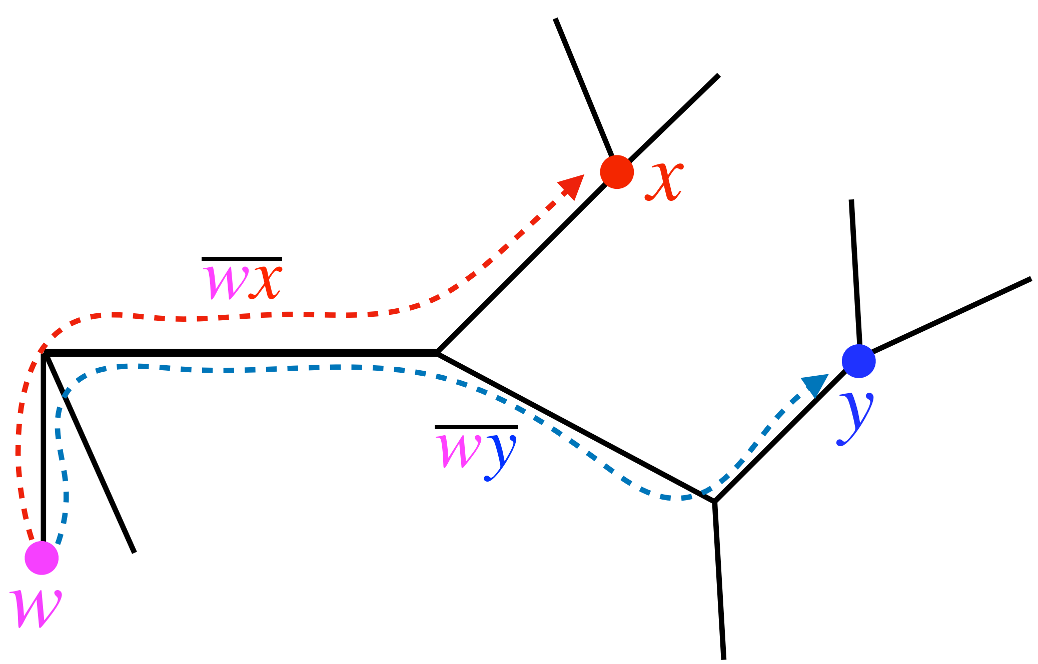
$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$

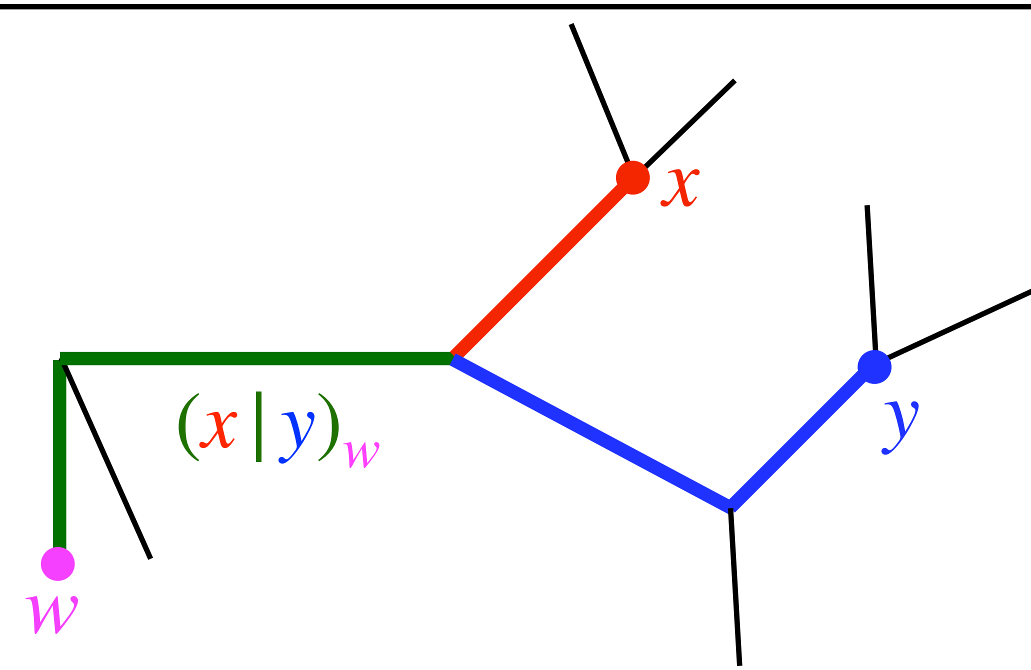


$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

✦ Interpretation

$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging

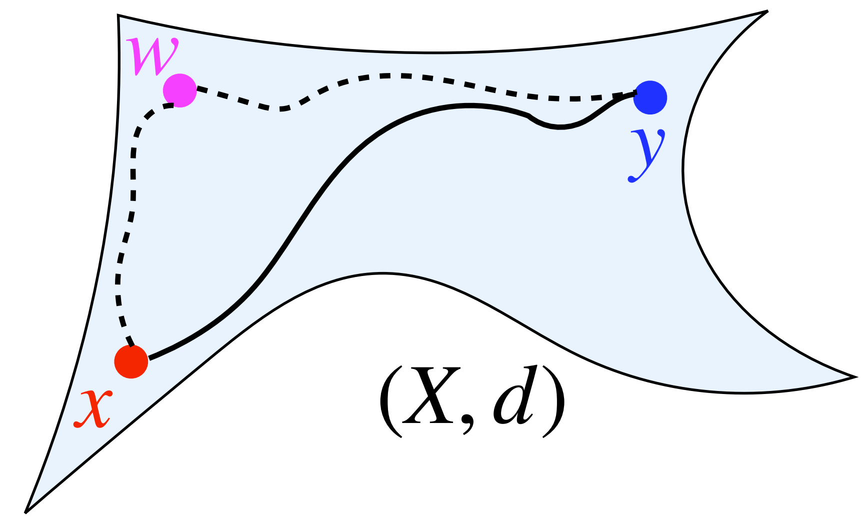




Gromov hyperbolicity

✦ Gromov product

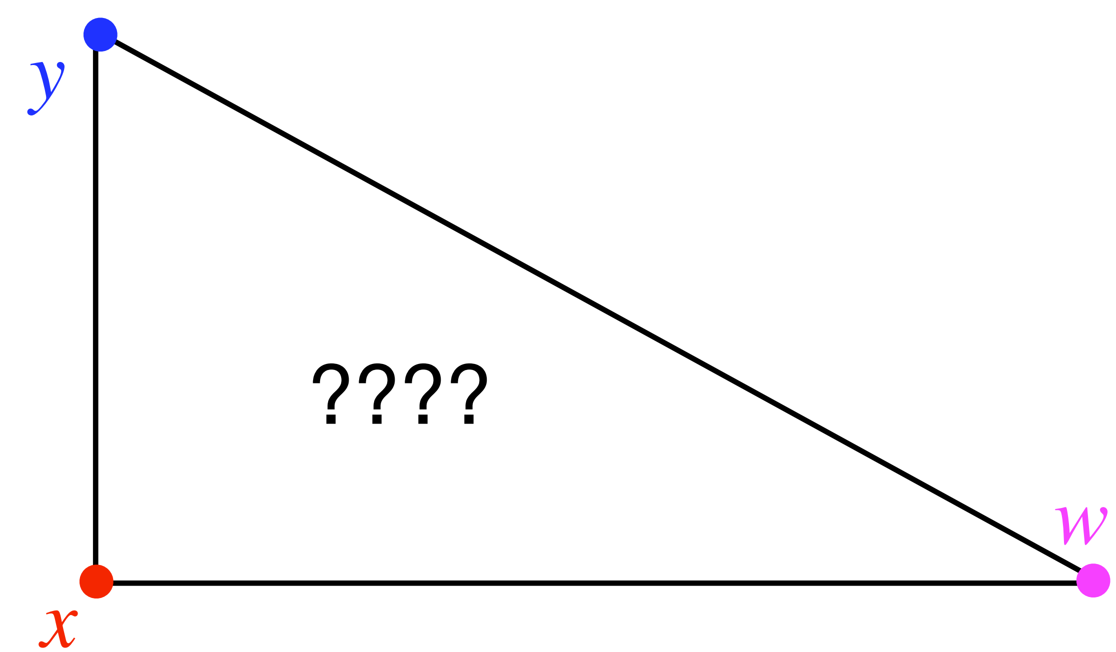
$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

✦ Interpretation

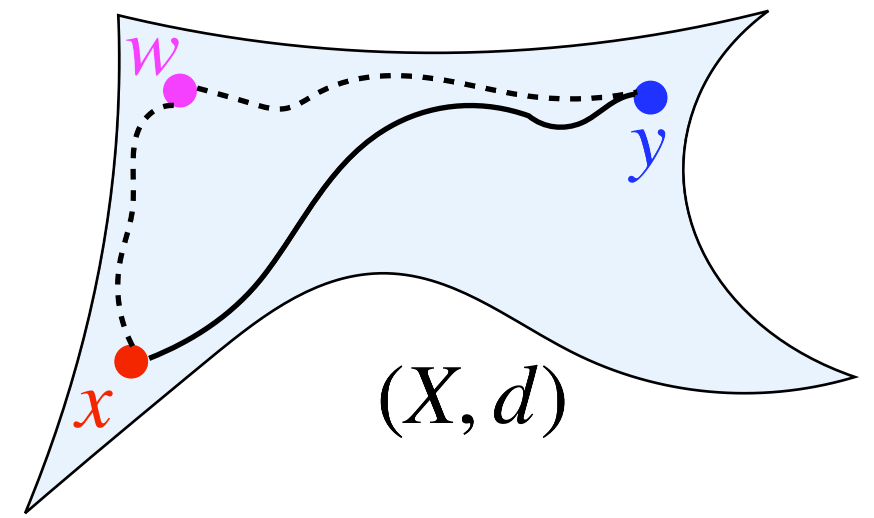
$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



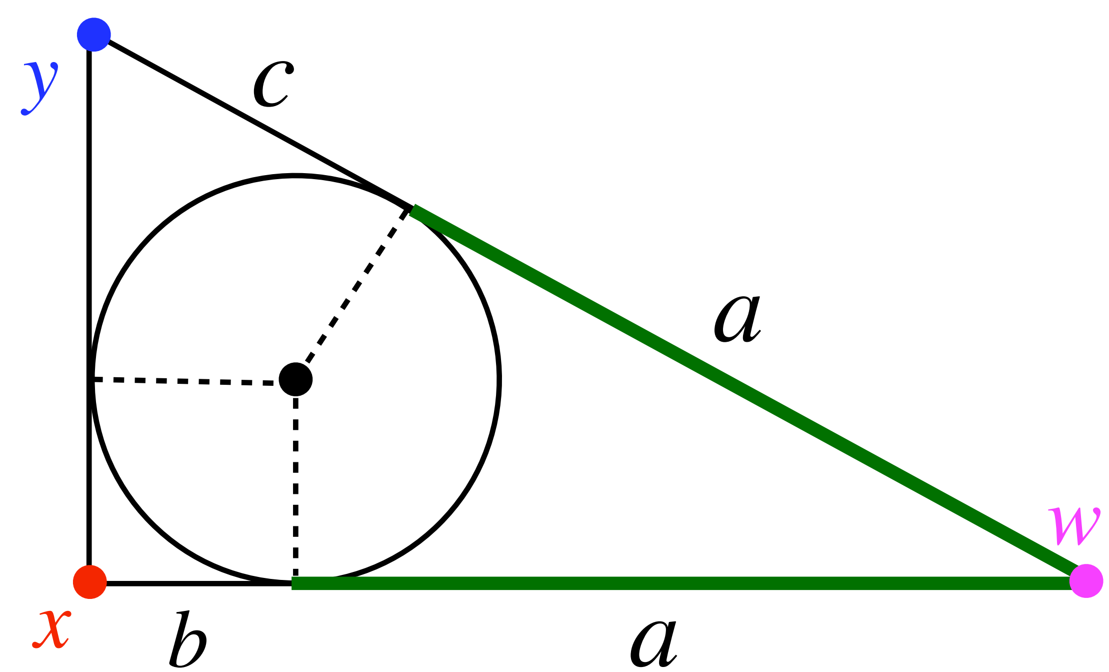
$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$

✦ Interpretation

$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging

$\exists!(a,b,c) \ge 0$

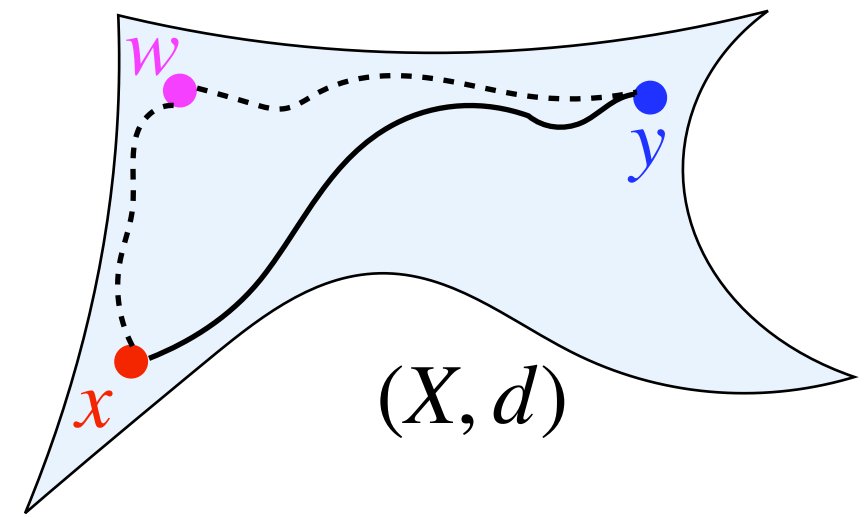
$$d(x,w) = a + b$$
$$d(x,y) = b + c$$
$$d(y,w) = a + c$$



Gromov hyperbolicity

✦ Gromov product

$$(x \mid y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$



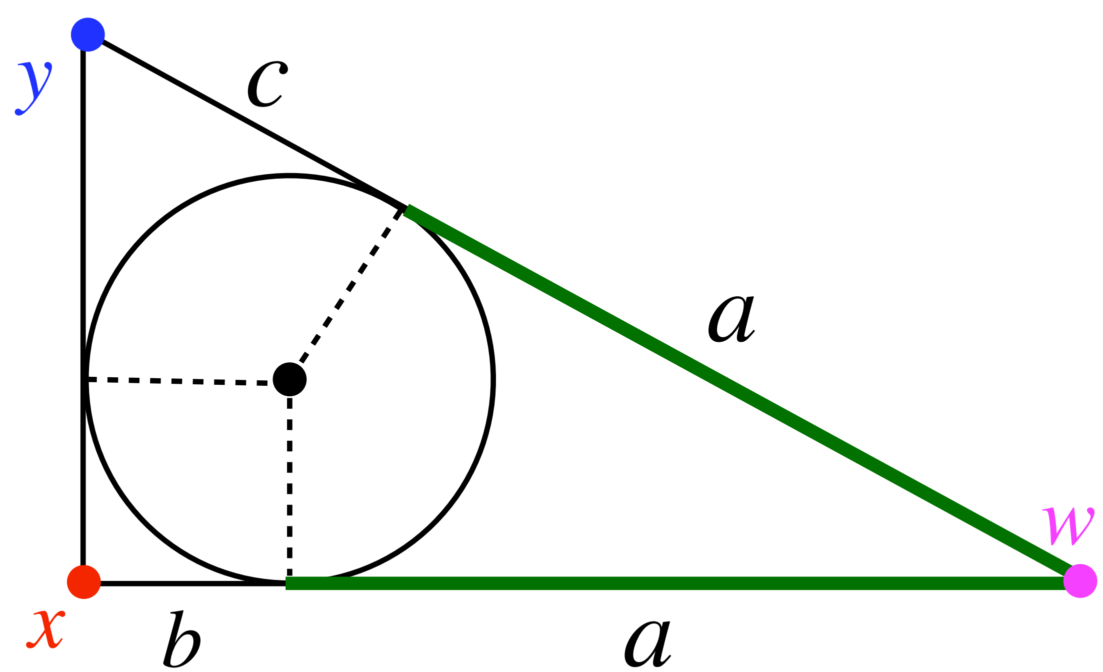
$$0 \leq (x \mid y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Interpretation

$(x \mid y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging

$\exists!(a,b,c) \geq 0$

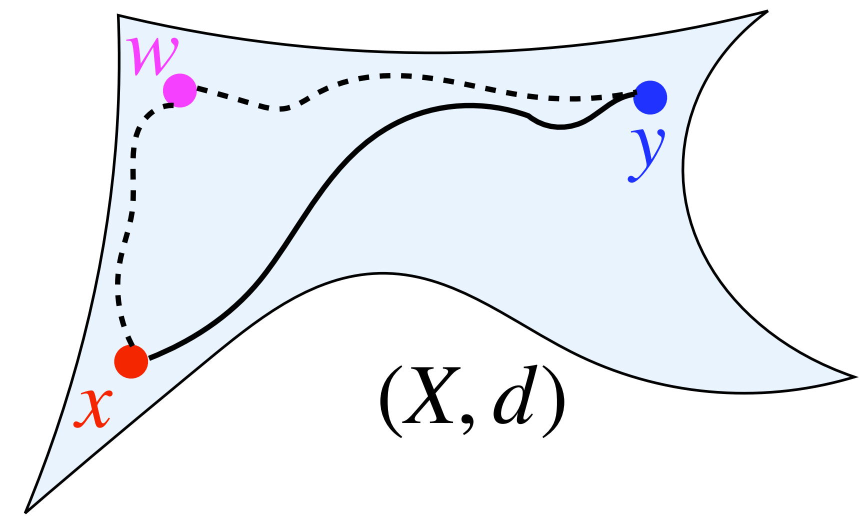
$$d(x,w) = a + b$$
$$d(x,y) = b + c \qquad \Longrightarrow \qquad a = (x \mid y)_w$$
$$d(y,w) = a + c$$



Gromov hyperbolicity

✦ Gromov product

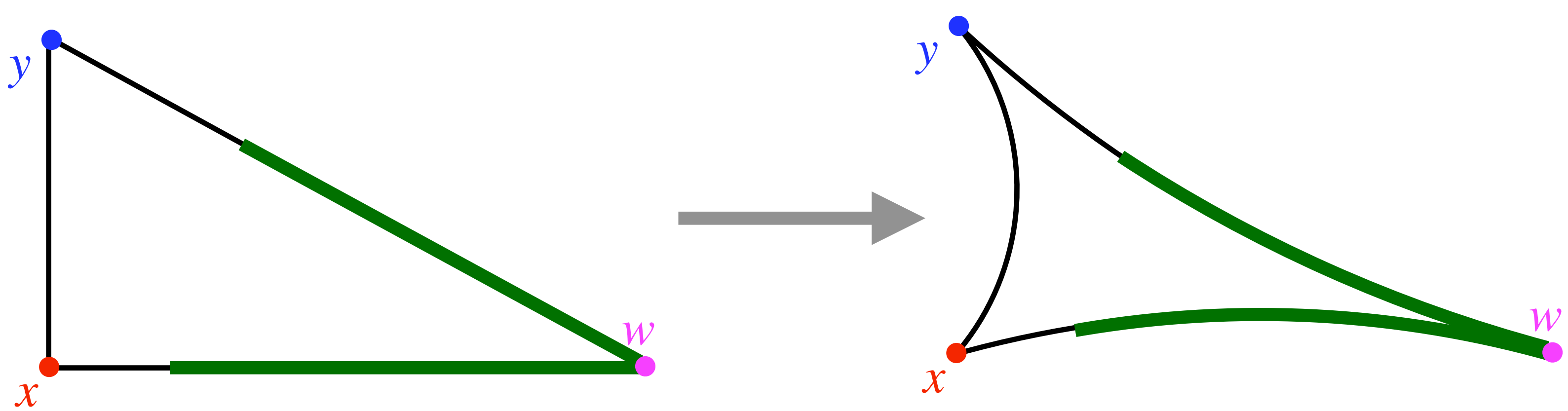
$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Interpretation

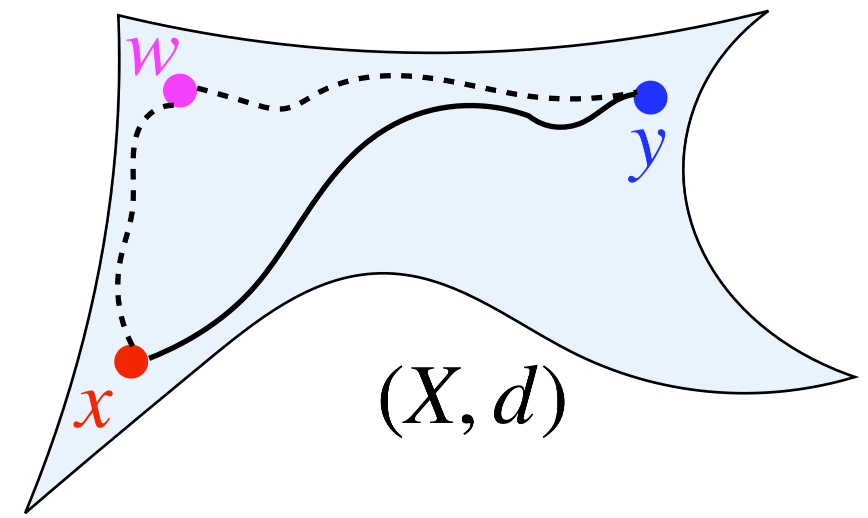
$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging



Gromov hyperbolicity

✦ Gromov product

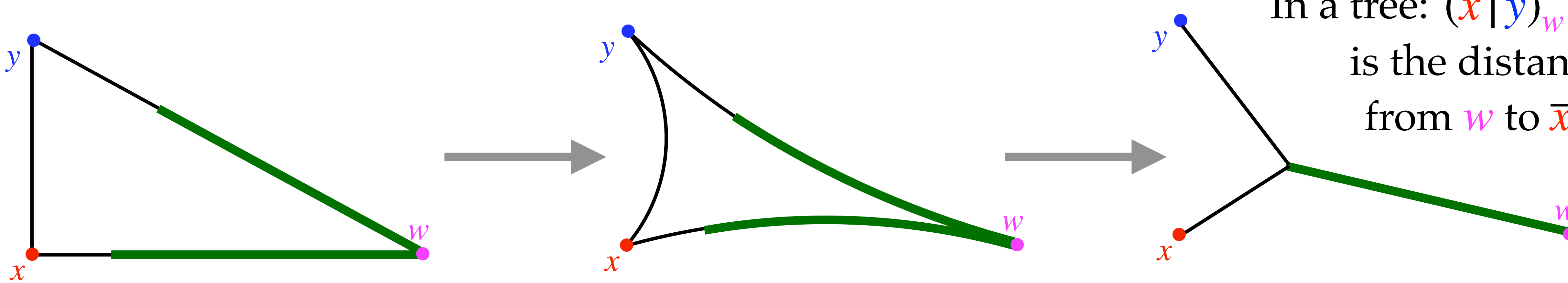
$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$

✦ Interpretation

$(x\,|\,y)_w$ measures how long geodesics $\overline{wx}$ and $\overline{wy}$ travel the same distance before diverging

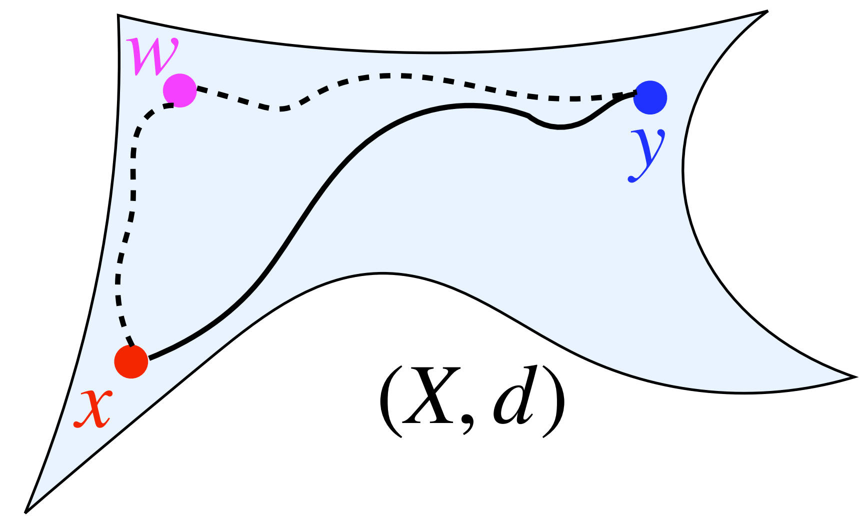


In a tree: $(x\,|\,y)_w$ is the distance from w to $\overline{xy}$

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Gromov δ-hyperbolicity

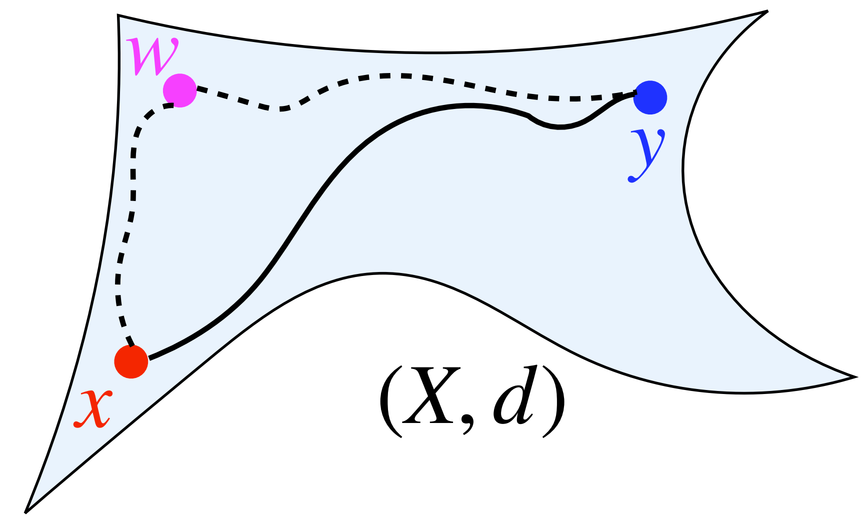
(X,d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x,y,z,w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

Gromov hyperbolicity

✦ Gromov product

$$(x \mid y)_w = \frac{1}{2}\Big(d(x, w) + d(y, w) - d(x, y)\Big)$$



$$0 \leq (x \mid y)_w \leq \min\{d(x, w), d(y, w)\}$$

✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x \mid y)_w \geq \min\{(y \mid z)_w\ ,\ (x \mid z)_w\} - \delta$$

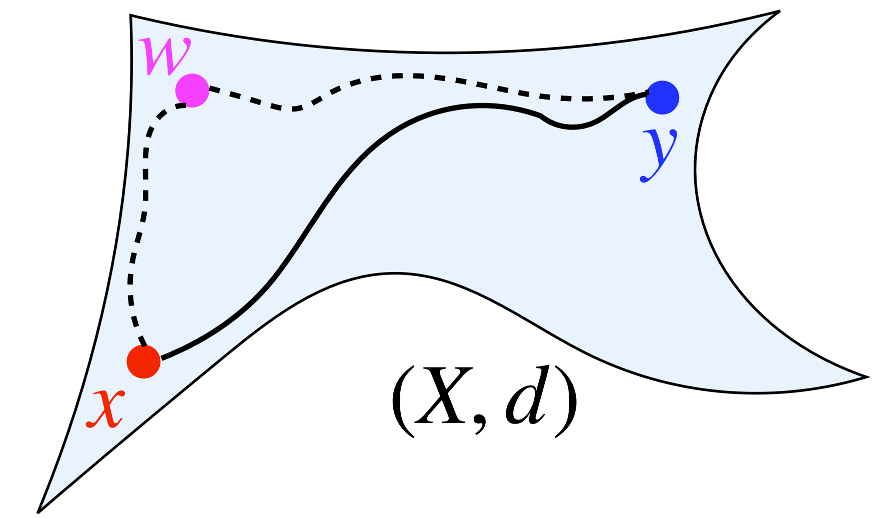
✦ Interpretation: δ-slim triangles

(X, d) is δ-hyperbolic iff each side of any Δ is contained within the δ -neighborhood of the union of the two other sides

Gromov hyperbolicity

✦ Gromov product

$$(x \mid y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x \mid y)_w \leq \min\{d(x,w), d(y,w)\}$$

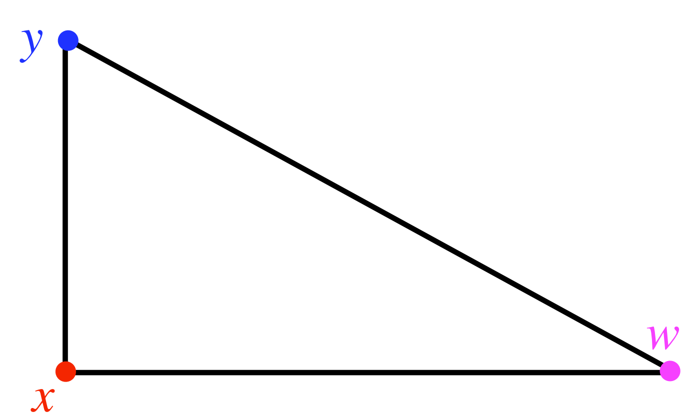
✦ Gromov δ-hyperbolicity

(X,d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x,y,z,w \in X,\ (x \mid y)_w \geq \min\{(y \mid z)_w\ ,\ (x \mid z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

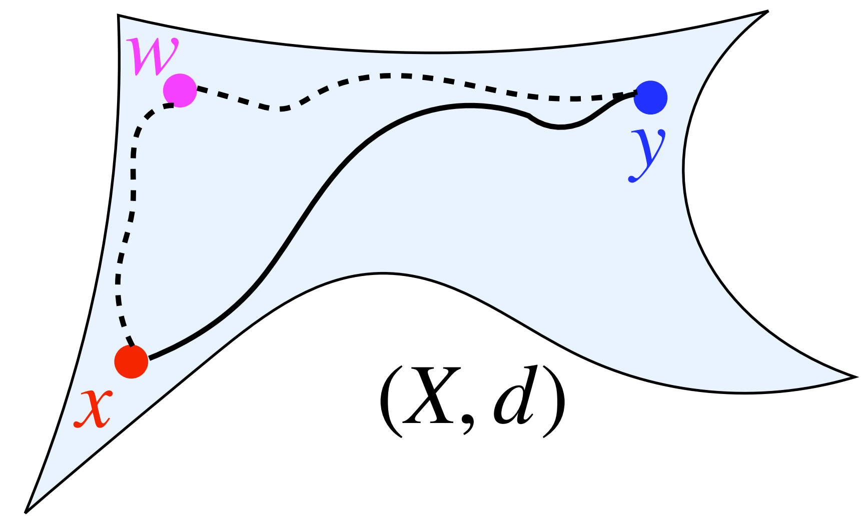
(X,d) is δ-hyperbolic iff each side of any Δ is contained within the δ -neighborhood of the union of the two other sides



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

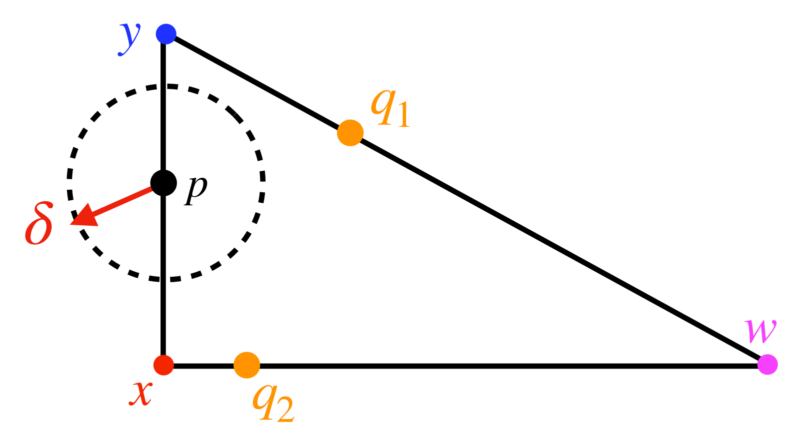
✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

$$\forall p \in [x,y],\ \exists q \in [y,w] \cup [x,w],\ d(p,q) \leq \delta$$

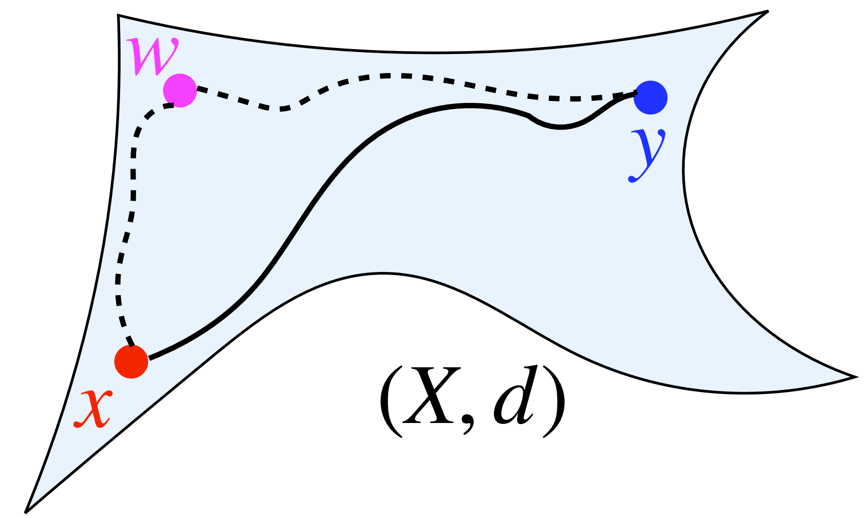


(X, d) is δ-hyperbolic iff each side of any Δ is contained within the δ-neighborhood of the union of the two other sides

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

✦ Gromov δ-hyperbolicity

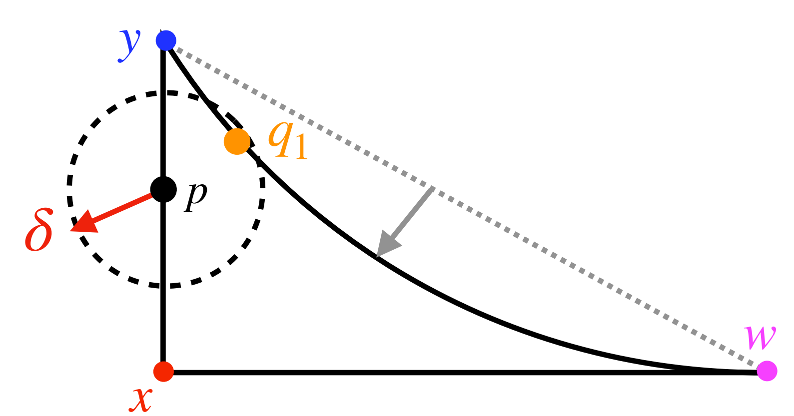
(X, d) is δ-hyperbolic (for $\delta \ge 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \ge \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

$$\forall p \in [x, y],\ \exists q \in [y, w] \cup [x, w],\ d(p, q) \le \delta$$

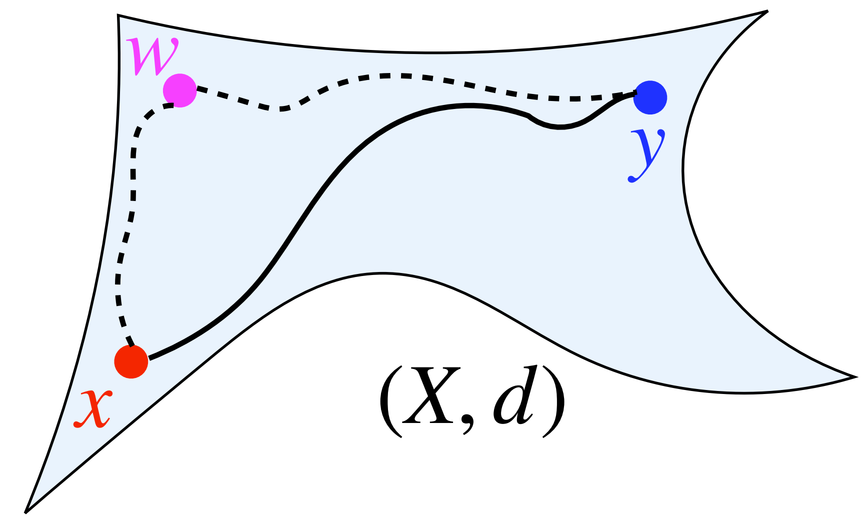
(X, d) is δ-hyperbolic iff each side of any Δ is contained within the δ -neighborhood of the union of the two other sides



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

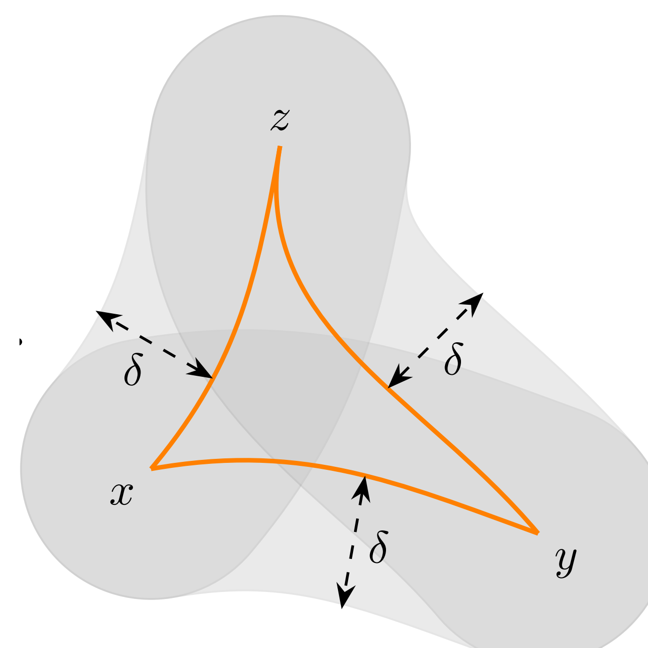
✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \ge 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \ge \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

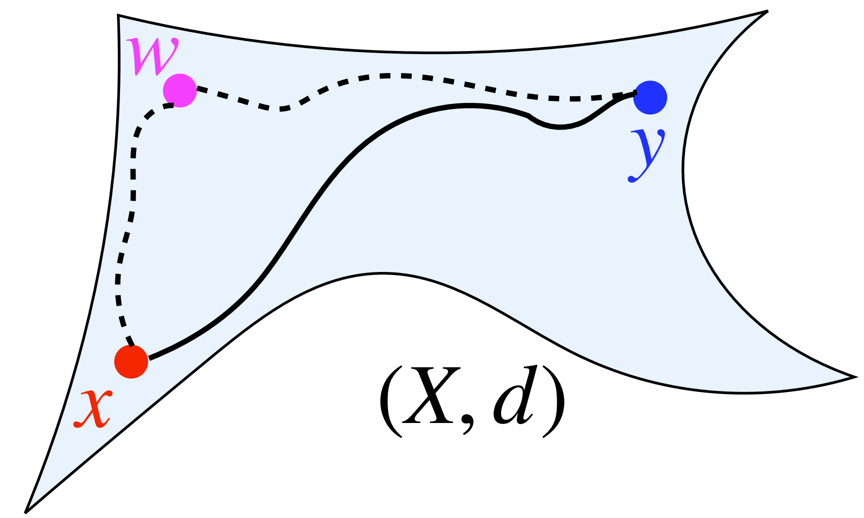
(X, d) is δ-hyperbolic iff each side of any Δ is contained within the δ -neighborhood of the union of the two other sides



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Gromov δ-hyperbolicity

(X,d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x,y,z,w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

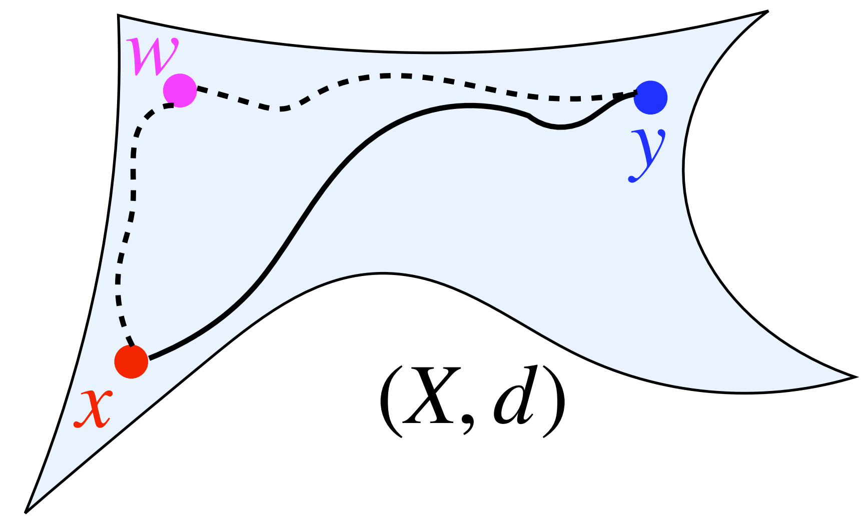
✦ Interpretation: δ-slim triangles

- ✦ Triangles in δ-hyperbolic spaces are « thin »
- ✦ **The smaller δ the thinner**

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$

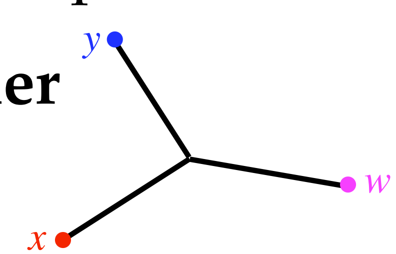
✦ Gromov δ-hyperbolicity

(X,d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x,y,z,w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

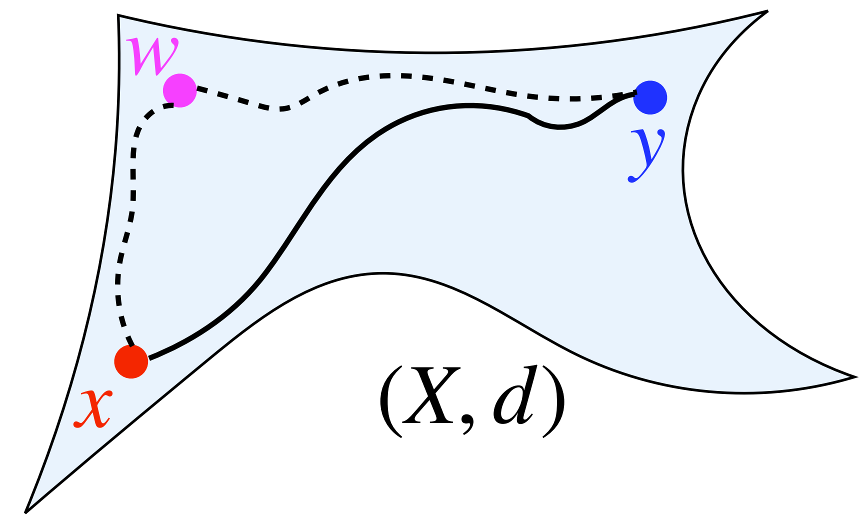
- Triangles in δ-hyperbolic spaces are « thin »
- **The smaller δ the thinner**
- Trees are 0-hyperbolic



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

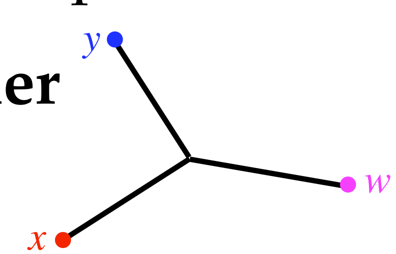
✦ Gromov δ-hyperbolicity

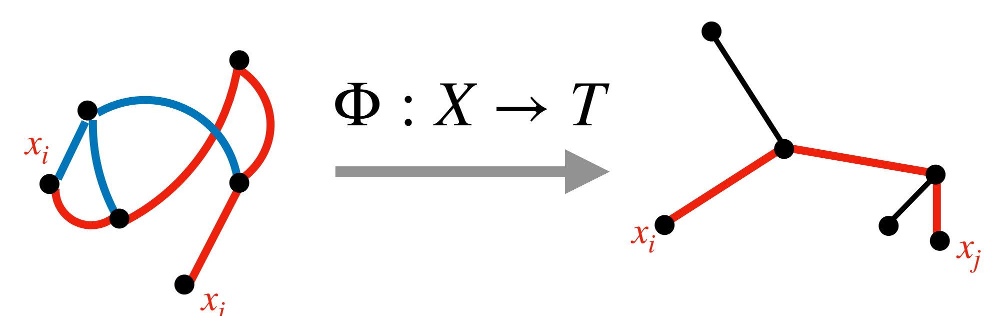
(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Interpretation: δ-slim triangles

- Triangles in δ-hyperbolic spaces are « thin »
- **The smaller δ the thinner**
- Trees are 0-hyperbolic
- If (X, d) is 0-hyperbolic **it is isometric to a tree**



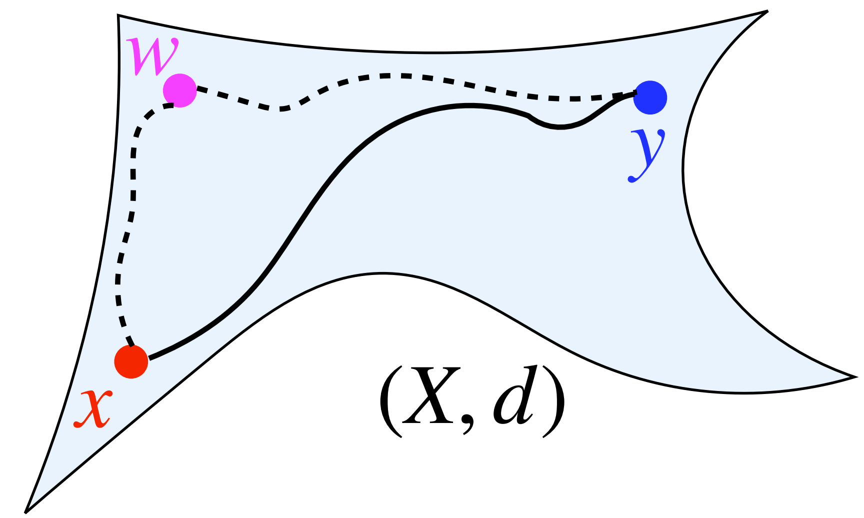


$$d(x_i, x_j) = d_T(\Phi(x_i), \Phi(x_j))$$

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



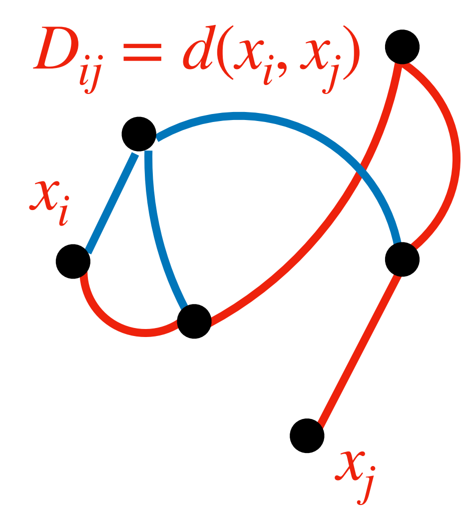
$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \ge 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \ge \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

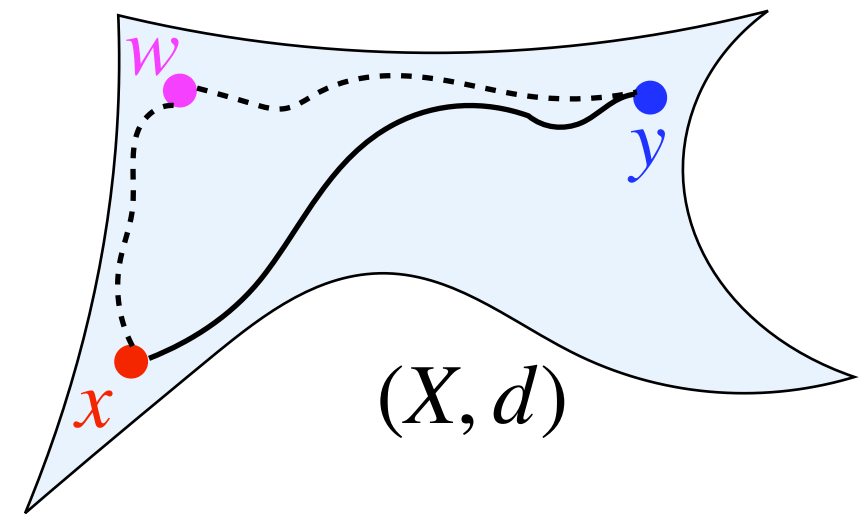
✦ Finite metric space $X = \{x_1, \cdots, x_n\}$



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

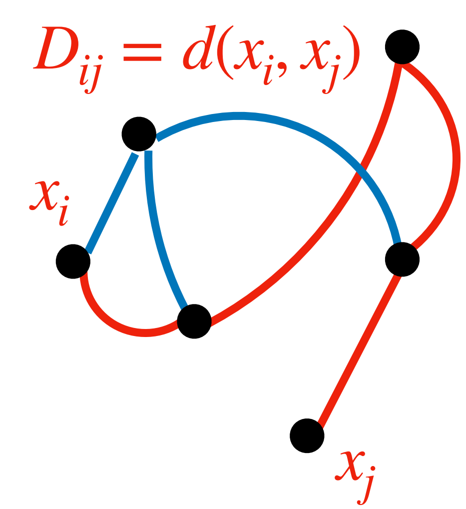
✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Finite metric space $X = \{x_1, \cdots, x_n\}$

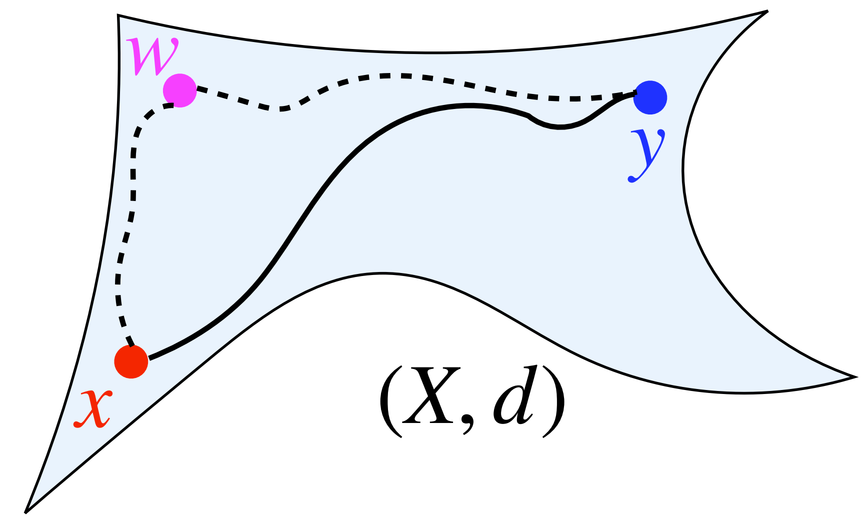
$$\delta_X = \max_{x,y,z,w}\Big(\min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - (x\,|\,y)_w\Big)$$



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \le (x\,|\,y)_w \le \min\{d(x,w), d(y,w)\}$$

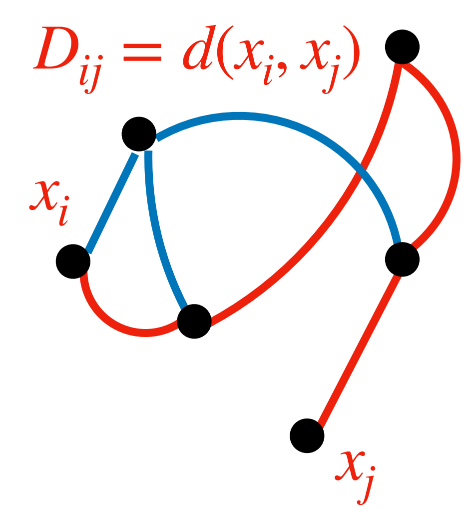
✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \ge 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \ge \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Finite metric space $X = \{x_1, \cdots, x_n\}$

$$\delta_X = \max_{x,y,z,w}\big(\min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - (x\,|\,y)_w\big)$$

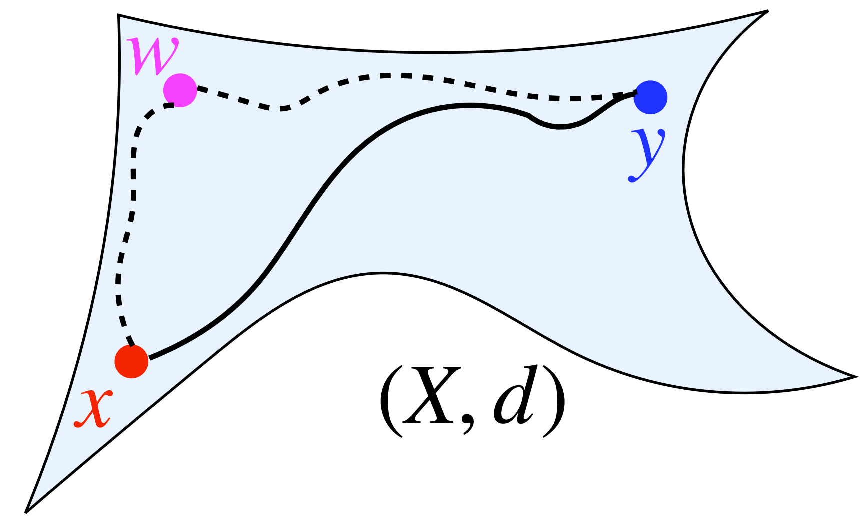


- ✦ Take $z = w = y$ this implies $\delta_X \ge 0$
- ✦ Computation in $O(n^4)$
- ✦ Depends only on d, we note $\delta_d, \delta_D \ldots$

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\big(d(x,w) + d(y,w) - d(x,y)\big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x,w), d(y,w)\}$$

✦ Gromov δ-hyperbolicity

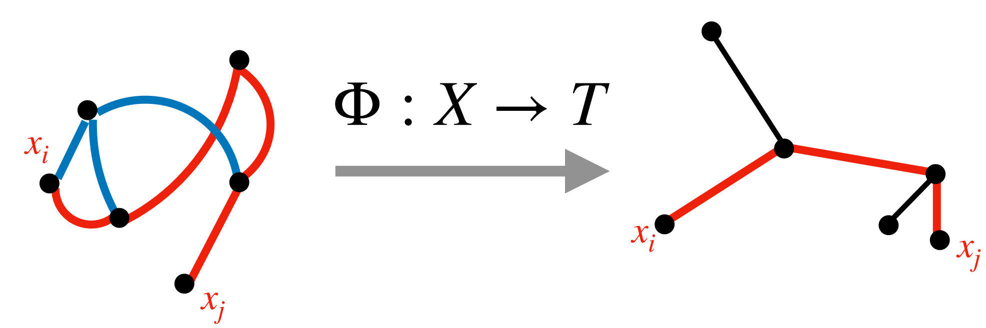
(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Finite metric space $X = \{x_1, \cdots, x_n\}$

$\exists T$ a tree, $\Phi : X \to T$

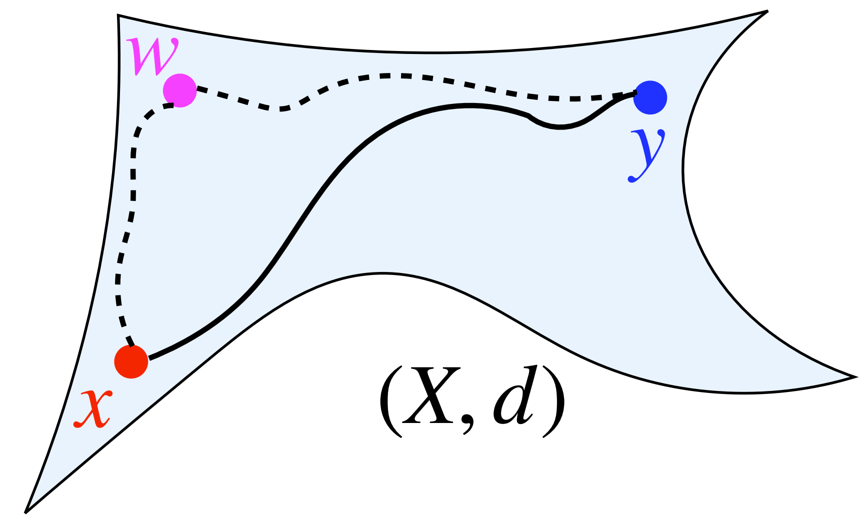
$$d(x_i, x_j) - 2\delta_X \log(n-2) \leq d_T(\Phi(x_i), \Phi(x_j)) \leq d(x_i, x_j)$$



Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x, w) + d(y, w) - d(x, y)\Big)$$



$$0 \leq (x\,|\,y)_w \leq \min\{d(x, w), d(y, w)\}$$

✦ Gromov δ-hyperbolicity

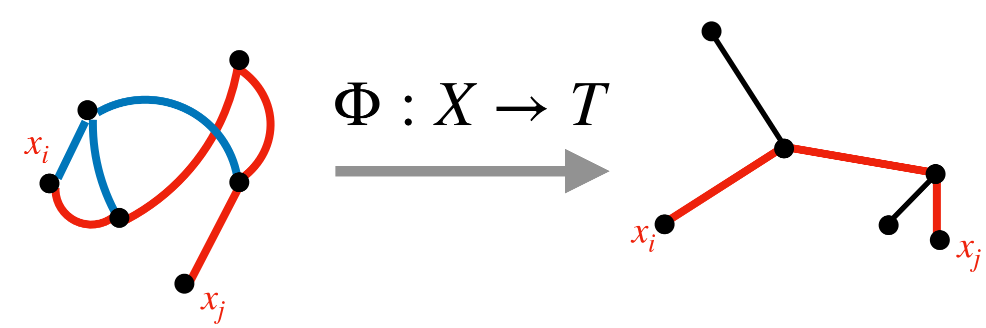
(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Finite metric space $X = \{x_1, \cdots, x_n\}$

$\exists T$ a tree, $\Phi : X \to T$

$$d(x_i, x_j) - 2\delta_X \log(n-2) \leq d_T(\Phi(x_i), \Phi(x_j)) \leq d(x_i, x_j)$$

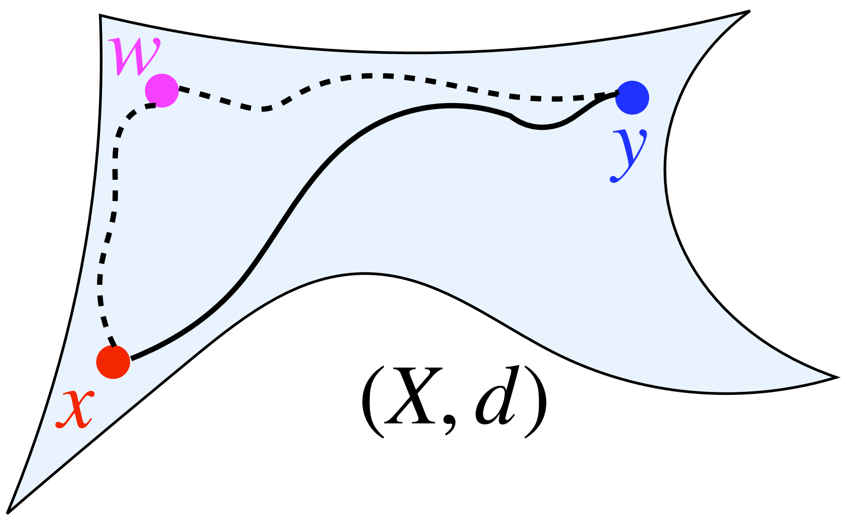


- ✦ It can be computed in $O(n^2)$
- ✦ Single Linkage Hierarchical Clustering algorithm
- ✦ This is the **Gromov embedding** Φ, $T = \text{Gromov}(X, d)$

Gromov hyperbolicity

✦ Gromov product

$$(x\,|\,y)_w = \frac{1}{2}\Big(d(x,w) + d(y,w) - d(x,y)\Big)$$



✦ Gromov δ-hyperbolicity

(X, d) is δ-hyperbolic (for $\delta \geq 0$) if

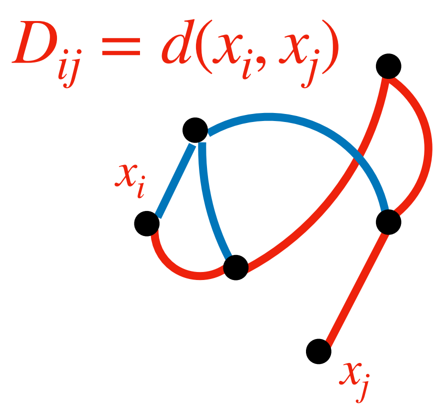
$$\forall x, y, z, w \in X,\ (x\,|\,y)_w \geq \min\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - \delta$$

✦ Conclusion

- ✦ δ-hyperbolicity quantifies to which extent (X, d) is hyperbolic
- ✦ The smaller δ the more (X, d) has a « tree structure »

To DeltaZero

✦ Motivation

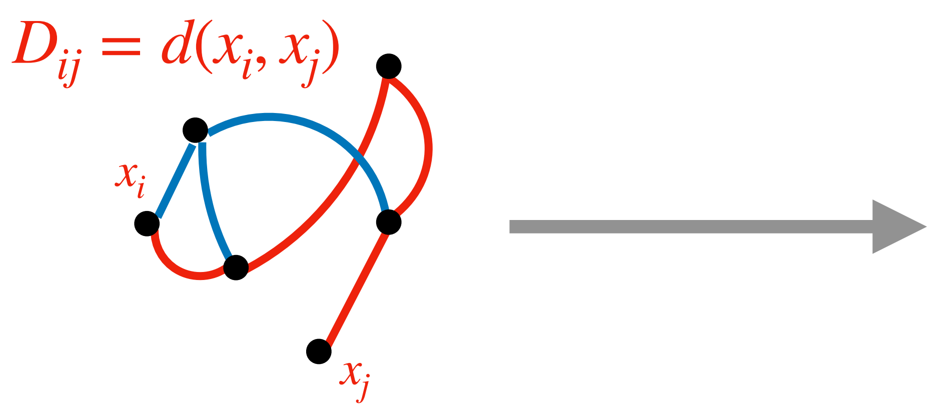


Embedding with a hierarchical structure

- ✦ Control how much « tree structure » we want
- ✦ Faithful to the original metric
- ✦ Reasonable in terms of computation

To DeltaZero

✦ Motivation



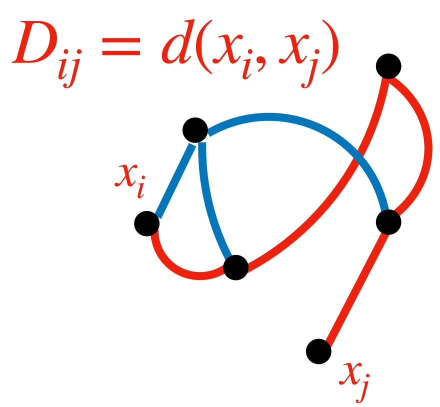
Embedding with a hierarchical structure

- ✦ Control how much « tree structure » we want
- ✦ Faithful to the original metric
- ✦ Reasonable in terms of computation

✦Observation: $$d^\star \in \underset{\substack{d' \in M_n \\ \forall i,j \; d'(x_i, x_j) \leq d(x_i, x_j)}}{\operatorname{argmin}} \mu \|d - d'\|_\infty + \delta_{d'}$$

To DeltaZero

✦ Motivation



Embedding with a hierarchical structure

- ✦ Control how much « tree structure » we want
- ✦ Faithful to the original metric
- ✦ Reasonable in terms of computation

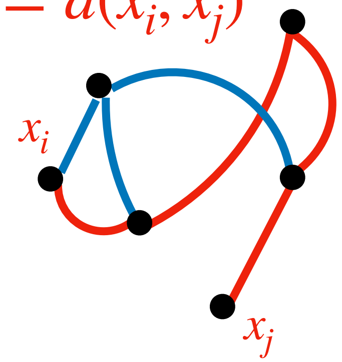
✦Observation:

$$d^\star \in \underset{\substack{d' \in M_n \\ \forall i,j\ d'(x_i,x_j) \leq d(x_i,x_j)}}{\operatorname{argmin}} \mu \|d - d'\|_\infty + \delta_{d'}$$

$$\Phi, T = \mathrm{Gromov}(X, d^\star)$$

To DeltaZero

✦ Motivation



Embedding with a hierarchical structure

- ✦ Control how much « tree structure » we want
- ✦ Faithful to the original metric
- ✦ Reasonable in terms of computation

✦Observation:

$$d^\star \in \underset{\substack{d' \in M_n \\ \forall i,j \; d'(x_i, x_j) \leq d(x_i, x_j)}}{\operatorname{argmin}} \mu \|d - d'\|_\infty + \delta_{d'}$$

$$\Phi, T = \mathrm{Gromov}(X, d^\star)$$

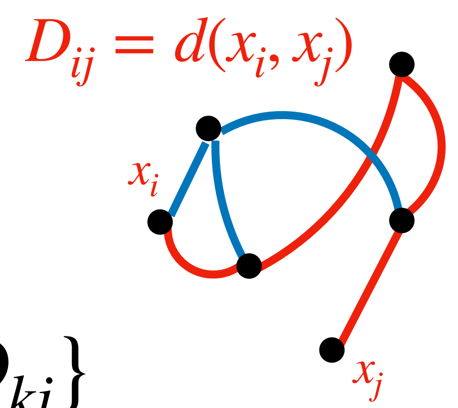
$$d(x_i, x_j) - 2\delta_X \log(n-2) + (2\log(n-2)\mu - 1)\|d - d^\star\|_\infty \leq d_T(\Phi(x_i), \Phi(x_j)) \leq d(x_i, x_j)$$

✦ When $\mu \geq 1/(2\log(n-2))$ we improve the lower bound

To DeltaZero

✦ Optimization problem

- ✦ Space of metrics on n points

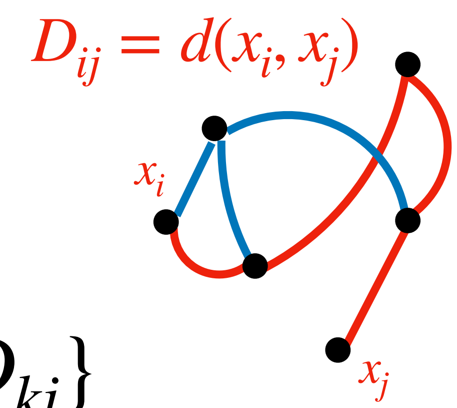


$$\mathscr{D}_n = \{D : \mathrm{diag}(D) = 0,\ D = D^\top,\ D_{ij} \leq D_{ik} + D_{kj}\}$$

$$\min_{\substack{D' \in \mathscr{D}_n \\ D' \leq D}} \; L(D) := \; \mu \|D - D'\|_F^2 + \delta_{D'}$$

To DeltaZero

✦ Optimization problem



- ✦ Space of metrics on n points

$$\mathscr{D}_n = \{D : \mathrm{diag}(D) = 0,\ D = D^\top,\ D_{ij} \le D_{ik} + D_{kj}\}$$

$$\min_{\substack{D' \in \mathscr{D}_n \\ D' \le D}} \ L(D) := \ \mu \|D - D'\|_F^2 + \delta_{D'}$$

- ✦ Trade-off between fidelity to D and small δ-hyperbolicity
- ✦ $D \to \delta_D$ is **piecewise affine** and **not convex**
- ✦ Complexity in $O(n^4)$ + not everywhere differentiable
- ✦ How to handle the constraints ?

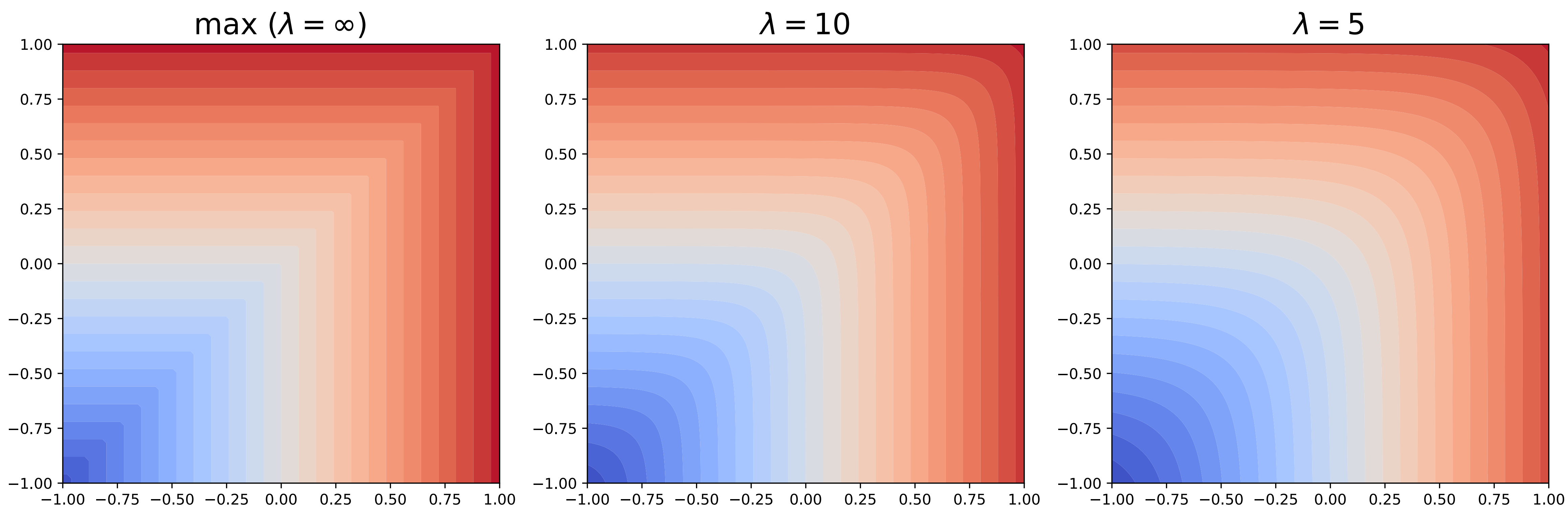
Smooth and batched δ-hyperbolicity

✦ Smoothing δ-hyperbolicity

- We replace the min , max in δ-hyperbolicity by a smooth surrogate

$$\mathrm{LSE}_\lambda(x) = \frac{1}{\lambda}\log\Big(\sum_i e^{\lambda x_i}\Big)$$

$$\max(x) \leq \mathrm{LSE}_\lambda(x) \leq \max(x) + \frac{\log(n)}{\lambda}$$



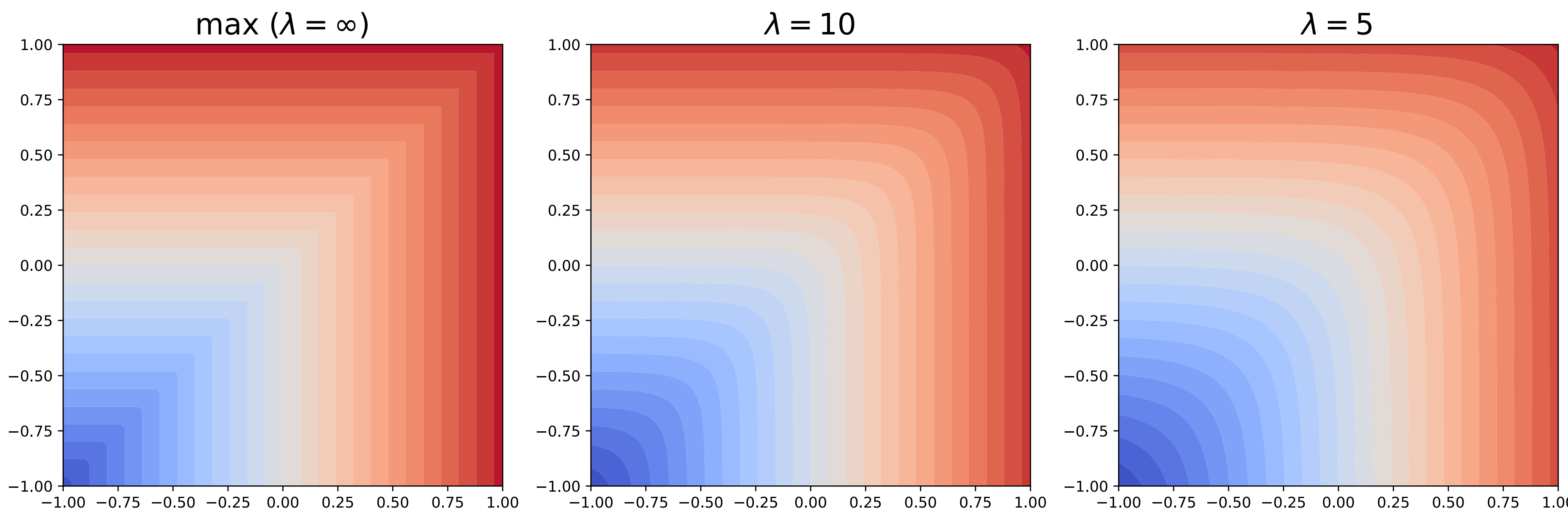
Smooth and batched δ-hyperbolicity

✦ Smoothing δ-hyperbolicity

- ✦ We replace the min , max in δ-hyperbolicity by a smooth surrogate

$$\mathrm{LSE}_\lambda(x) = \frac{1}{\lambda}\log(\sum_i e^{\lambda x_i})$$

$$\max(x) \leq \mathrm{LSE}_\lambda(x) \leq \max(x) + \frac{\log(n)}{\lambda}$$



- ✦ Corresponds to regularizing a linear problem

$$\max(x) = \max_{q \in \Sigma_n} \langle x, q \rangle \longrightarrow \mathrm{LSE}_\lambda(x) = \operatorname{argmax}_{q \in \Sigma_n} \langle x, q \rangle - \frac{1}{\lambda}\mathrm{entropy}(q)$$

Smooth and batched δ-hyperbolicity

✦ Smoothing δ-hyperbolicity

- We replace the min , max in δ-hyperbolicity by a smooth surrogate

$$\mathrm{LSE}_\lambda(x) = \frac{1}{\lambda}\log\Big(\sum_i e^{\lambda x_i}\Big)$$

- Smoothed δ-hyperbolicity: differentiable but still $O(n^4)$

$$\delta^{(\lambda)} = \mathrm{LSE}_\lambda\left(\Big\{\mathrm{LSE}_{-\lambda}\{(y\,|\,z)_w\ ,\ (x\,|\,z)_w\} - (x\,|\,y)_w\Big\}_{x,y,z,w}\right)$$

$$\delta - \frac{\log(2)}{\lambda} \le \delta^{(\lambda)} \le \delta + \frac{4\log(n)}{\lambda}$$

Smooth and batched δ-hyperbolicity

Smoothing δ-hyperbolicity

- We replace the min , max in δ-hyperbolicity by a smooth surrogate

$$\mathrm{LSE}_\lambda(x) = \frac{1}{\lambda}\log\Big(\sum_i e^{\lambda x_i}\Big)$$

- Smoothed δ-hyperbolicity: differentiable but still $O(n^4)$

$$\delta^{(\lambda)} = \mathrm{LSE}_\lambda\left(\left\{\mathrm{LSE}_{-\lambda}\{(y|z)_w\ ,\ (x|z)_w\} - (x|y)_w\right\}_{x,y,z,w}\right)$$

Batched δ-hyperbolicity

- Sample m points among the n
- Do that K times: gives $X_1, \cdots, X_K \subset X$
- Compute $\delta^{(\lambda)}_{X_1}, \cdots, \delta^{(\lambda)}_{X_K}$
- Complexity $O(K \cdot m^4)$

Smooth and batched δ-hyperbolicity

Smoothing δ-hyperbolicity

- We replace the min , max in δ-hyperbolicity by a smooth surrogate

$$\mathrm{LSE}_\lambda(x) = \frac{1}{\lambda}\log\Big(\sum_i e^{\lambda x_i}\Big)$$

- Smoothed δ-hyperbolicity: differentiable but still $O(n^4)$

$$\delta^{(\lambda)} = \mathrm{LSE}_\lambda\Big(\big\{\mathrm{LSE}_{-\lambda}\{(y|z)_w\ ,\ (x|z)_w\} - (x|y)_w\big\}_{x,y,z,w}\Big)$$

Batched δ-hyperbolicity

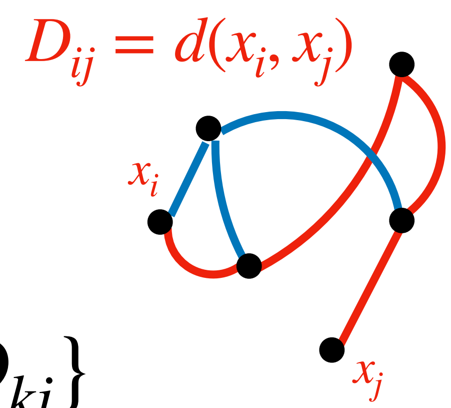
- Sample m points among the n
- Do that K times: gives $X_1, \cdots, X_K \subset X$
- Compute $\delta^{(\lambda)}_{X_1}, \cdots, \delta^{(\lambda)}_{X_K}$
- Complexity $O(K \cdot m^4)$

$$\delta^{(\lambda)}_{K,m} = \mathrm{LSE}_\lambda(\delta^{(\lambda)}_{X_1}, \cdots, \delta^{(\lambda)}_{X_K})$$

- Under some hypothesis, close to δ with high prob.

DeltaZero

✦ Optimization problem



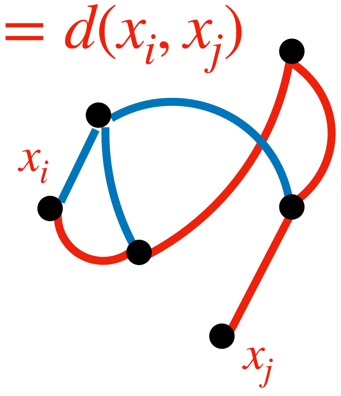
- ✦ Space of metrics on n points

$$\mathscr{D}_n = \{D : \mathrm{diag}(D) = 0,\ D = D^\top,\ D_{ij} \leq D_{ik} + D_{kj}\}$$

$$\min_{\substack{D' \in \mathscr{D}_n \\ D' \leq D}} L(D) := \ \mu \|D - D'\|_F^2 + \delta^{(\lambda)}_{D',K,m}$$

DeltaZero

✦ Optimization problem



- Space of metrics on n points

$$\mathscr{D}_n = \{D : \mathrm{diag}(D) = 0,\ D = D^\top,\ D_{ij} \le D_{ik} + D_{kj}\}$$

$$\min_{\substack{D' \in \mathscr{D}_n \\ D' \le D}} \ L(D) := \ \mu \|D - D'\|_F^2 + \delta^{(\lambda)}_{D',K,m}$$

- Projected gradient descent

Algorithm

$G_t = \nabla L(D_t)$

$D_{t+\frac{1}{2}} = \mathrm{Adam}(G_t, D_t)$

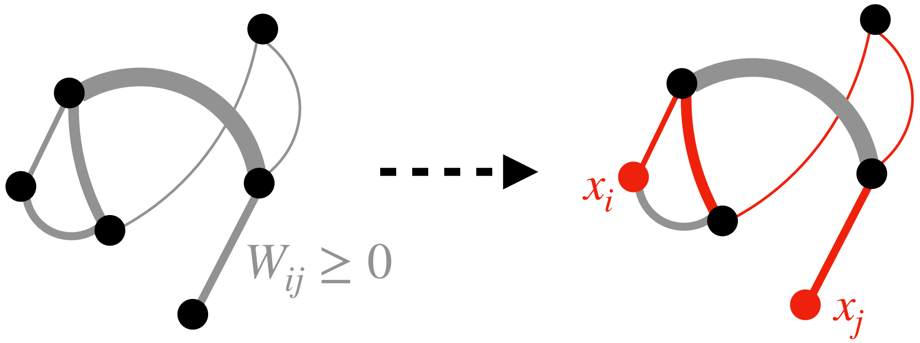
$D_t = \Pi(D_{t+\frac{1}{2}})$

$$\Pi(W) = \operatorname*{argmin}_{D \in \mathscr{D}_n : D \le W} \|D - W\|_F^2$$

- How to compute this projection ?
- Answer: this is a **shortest path problem**

DeltaZero

✦ The metric nearest problem



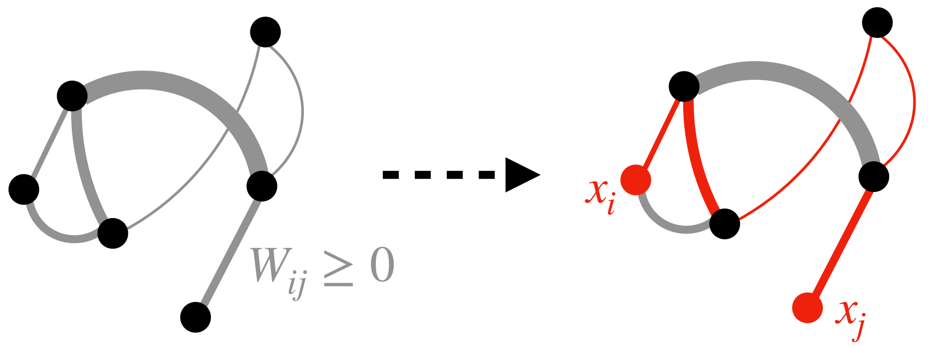
$$D = \text{ShortestPath}(W) = \begin{bmatrix} 0 & & D_{ij} \\ & \ddots & \\ & & 0 \end{bmatrix}$$

$\forall (i, j)$:

- ✦ Case 1) $D_{ij} = W_{ij}$
- ✦ Case 2) $\exists k, D_{ij} = D_{ik} + D_{kj}$

DeltaZero

✦ The metric nearest problem



$$D = \text{ShortestPath}(W) = \begin{bmatrix} 0 & & D_{ij} \\ & \ddots & \\ & & 0 \end{bmatrix}$$

$\forall (i, j)$:

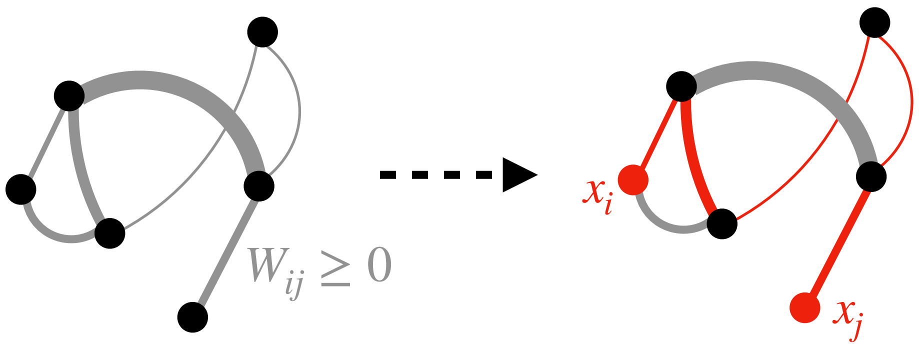
- ✦ Case 1) $D_{ij} = W_{ij}$
- ✦ Case 2) $\exists k, D_{ij} = D_{ik} + D_{kj}$

(Brickell, 2008)

$$\forall p, \ \underset{D \in \mathcal{D}_n : D \leq W}{\text{argmin}} \ \|D - W\|_{\ell_p} = \text{ShortestPath}(W)$$

DeltaZero

✦ The metric nearest problem



$$D = \text{ShortestPath}(W) = \begin{bmatrix} 0 & & D_{ij} \\ & \ddots & \\ & & 0 \end{bmatrix}$$

$\forall (i, j)$:

- ✦ Case 1) $D_{ij} = W_{ij}$
- ✦ Case 2) $\exists k, D_{ij} = D_{ik} + D_{kj}$

(Brickell, 2008)

$$\forall p, \quad \underset{D \in \mathscr{D}_n : D \leq W}{\text{argmin}} \; \|D - W\|_{\ell_p} = \text{ShortestPath}(W)$$

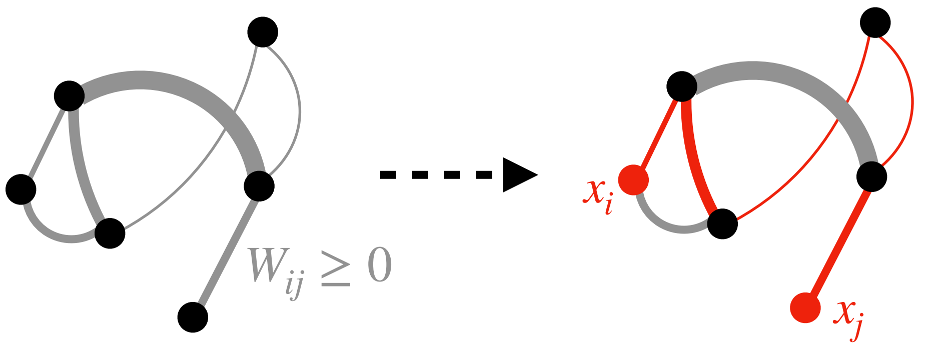
✦ Floyd-Warshall

```
D = W
for k ∈ {1,⋯,n}
    for i ∈ {1,⋯,n}
        for j ∈ {1,⋯,n}
            D_ij = min{D_ij, D_ik + D_kj}
```

- ✦ Finds ShortestPath(W)
- ✦ Runs in $O(n^3)$

DeltaZero

✦ The metric nearest problem



$$D = \text{ShortestPath}(W) = \begin{bmatrix} 0 & & D_{ij} \\ & \ddots & \\ & & 0 \end{bmatrix}$$

$\forall (i, j)$:

- ✦ Case 1) $D_{ij} = W_{ij}$
- ✦ Case 2) $\exists k, D_{ij} = D_{ik} + D_{kj}$

(Brickell, 2008)

$$\forall p, \quad \underset{D \in \mathcal{D}_n : D \leq W}{\text{argmin}} \quad \|D - W\|_{\ell_p} = \text{ShortestPath}(W)$$

✦ Floyd-Warshall

$D = W$
for $k \in \{1, \cdots, n\}$
  for $i \in \{1, \cdots, n\}$
    for $j \in \{1, \cdots, n\}$
      $D_{ij} = \min\{D_{ij}, D_{ik} + D_{kj}\}$

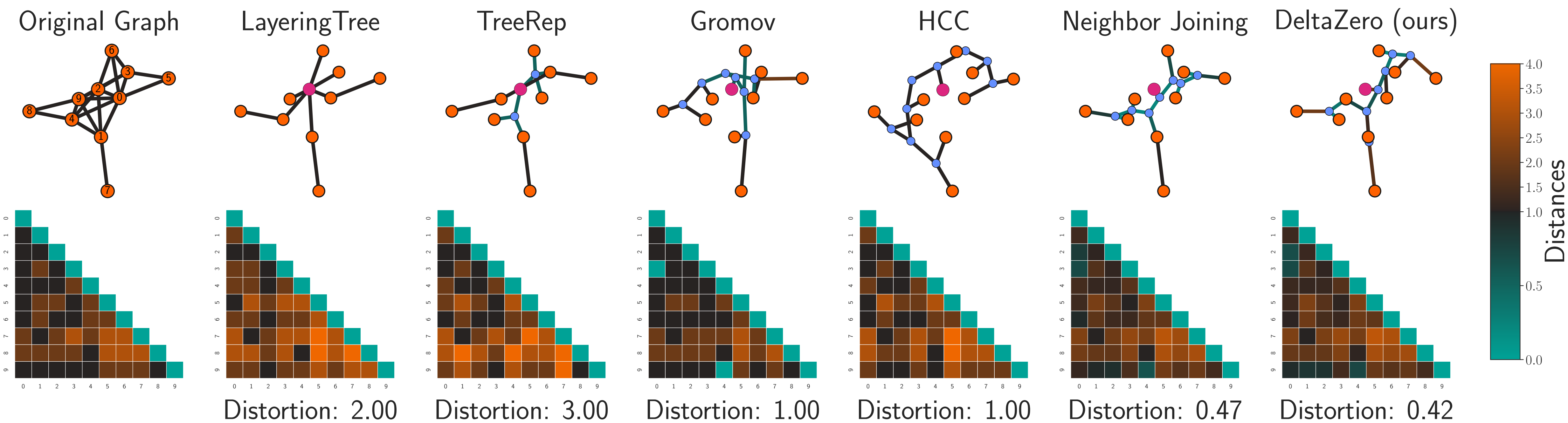
- ✦ Finds ShortestPath(W)
- ✦ Runs in $O(n^3)$

✦ DeltaZero

$G_t = \nabla L(D_t)$ // $O(K \cdot m^4 + n^2)$
$D_{t+\frac{1}{2}} = \text{Adam}(G_t, D_t)$
$D_t = \text{FloydWarshall}(D_{t+\frac{1}{2}})$ // $O(n^3)$
if output tree: $\Phi, T = \text{Gromov}(D_\infty)$ // $O(n^2)$

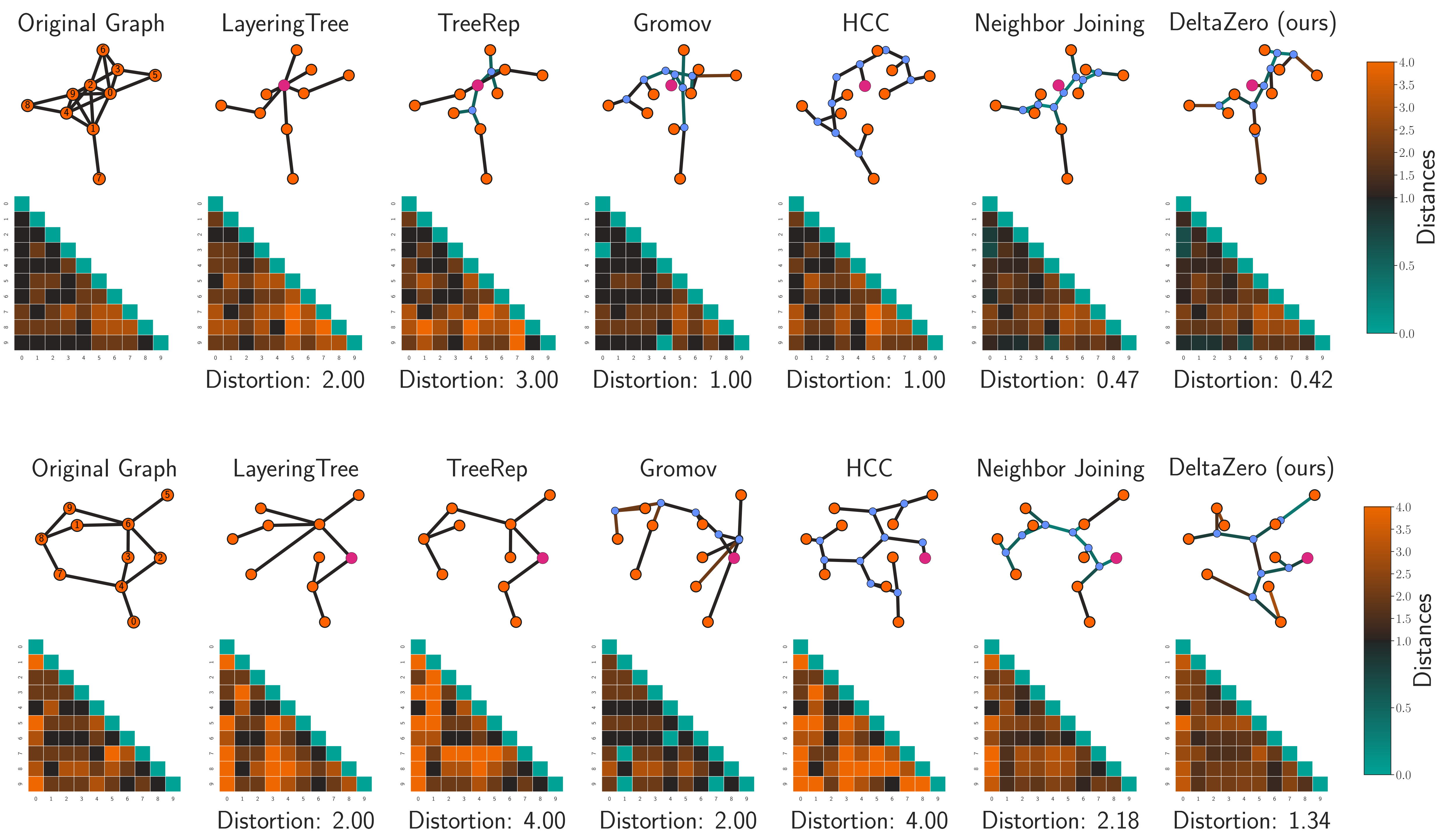
DeltaZero

✦ Ilustrations



DeltaZero

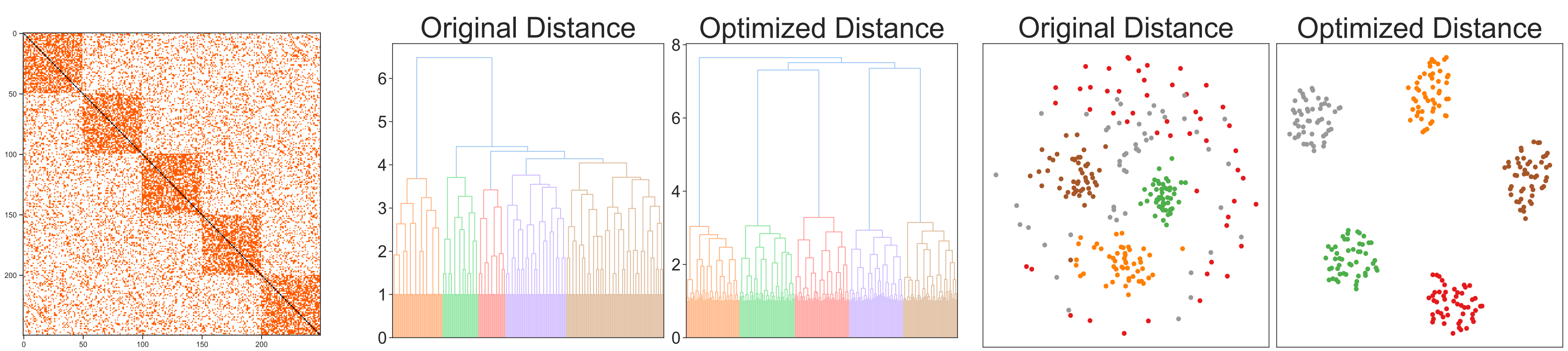
✦ Ilustrations



DeltaZero

✦ On stochastic block model

- SBM with 5 communities. Gives a shortest path matrix D
- Objective: clustering the nodes of the graph given D
- We compute $D' = \text{DeltaZero}(D)$
- We compare clustering (single linkage) with D vs D'



(a) All pairs Shortest-Paths distance matrix D

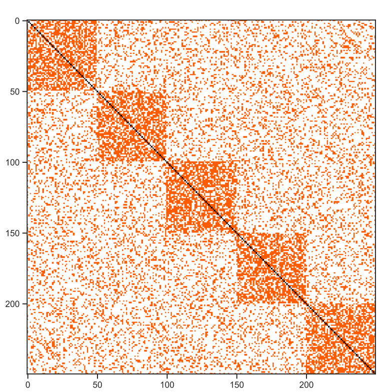
(b) Dendrograms from original and optimized distance matrices

(c) t-SNE plots from original and optimized distance matrices

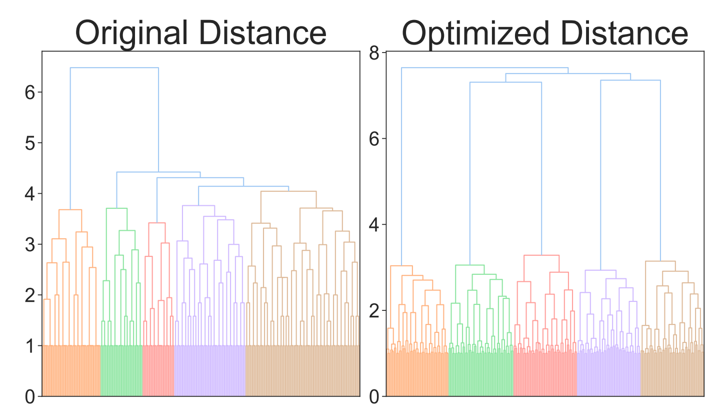
DeltaZero

✦ On stochastic block model

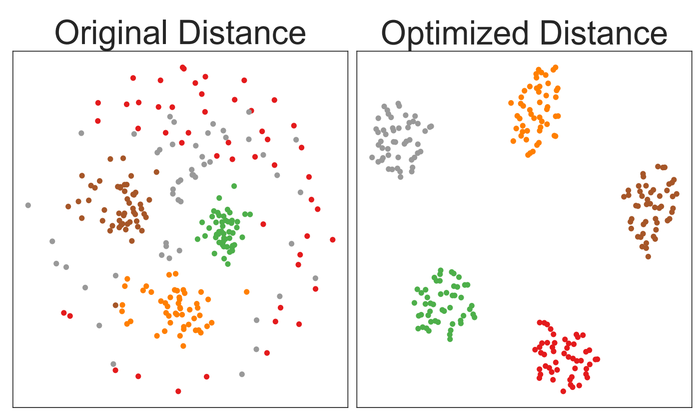
- SBM with 5 communities. Gives a shortest path matrix D
- Objective: clustering the nodes of the graph given D
- We compute $D' = \text{DeltaZero}(D)$
- We compare clustering (single linkage) with D vs D'



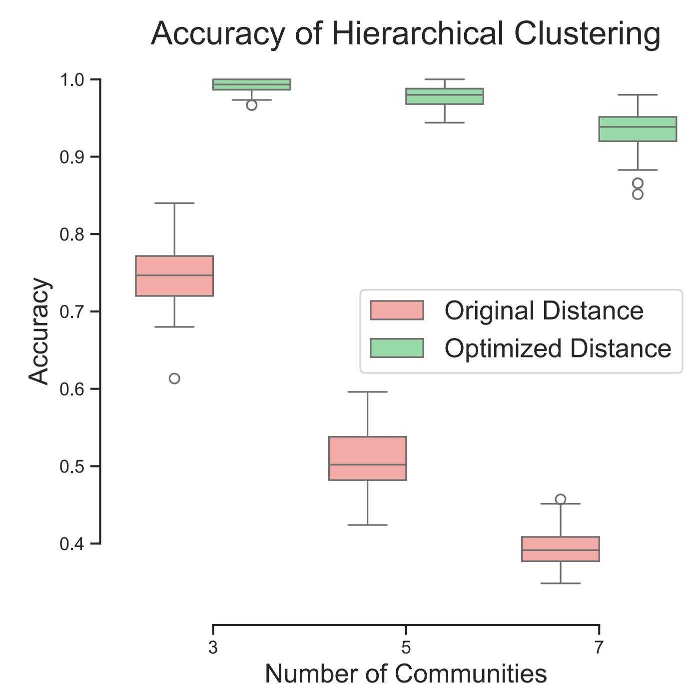
(a) All pairs Shortest-Paths distance matrix D



(b) Dendrograms from original and optimized distance matrices



(c) t-SNE plots from original and optimized distance matrices



DeltaZero

✦ Distorsion on real datasets

- ✦ We compute the tree metric D_T = DeltaZero(D) + Gromov
- ✦ We evaluate $\|D_T - D\|_\infty$

	Unweighted graphs					Non-graph metrics	
Datasets	C-ELEGAN	CS PHD	CORA	AIRPORT	WIKI	ZEISEL	IBD
n	452	1025	2485	3158	2357	3005	396
Diameter	7	28	19	12	9	0.87	0.99
NJ	<u>2.97</u>	16.81	13.42	4.18	6.32	0.51	<u>0.90</u>
TR	5.90 ± 0.72	21.01 ± 3.34	16.86 ± 2.11	10.00 ± 1.02	9.97 ± 0.93	0.66 ± 0.10	1.60 ± 0.22
HCC	4.31 ± 0.46	23.35 ± 2.07	12.28 ± 0.96	7.71 ± 0.72	7.20 ± 0.60	0.53 ± 0.07	1.25 ± 0.11
LayeringTree	5.07 ± 0.25	25.48 ± 0.60	<u>7.76</u> ± 0.54	<u>2.97</u> ± 0.26	<u>4.08</u> ± 0.27	–	–
Gromov	3.33 ± 0.45	<u>13.28</u> ± 0.61	9.34 ± 0.53	4.08 ± 0.27	5.54 ± 0.49	<u>0.43</u> ± 0.02	1.01 ± 0.04
DELTAZERO	**1.87** ± 0.08	**10.31** ± 0.62	**7.59** ± 0.38	**2.79** ± 0.15	**3.56** ± 0.20	**0.24** ± 0.00	**0.70** ± 0.03
Improvement (%)	43.8%	22.3%	2.3%	6.0%	12.7%	44.1 %	22.2%

DeltaZero

✦ Distorsion on real datasets

- ✦ We compute the tree metric $D_T = \mathrm{DeltaZero}(D) + \mathrm{Gromov}$
- ✦ We evaluate $\|D_T - D\|_\infty$

	Unweighted graphs					Non-graph metrics	
Datasets	C-ELEGAN	CS PHD	CORA	AIRPORT	WIKI	ZEISEL	IBD
n	452	1025	2485	3158	2357	3005	396
Diameter	7	28	19	12	9	0.87	0.99
NJ	2.97	16.81	13.42	4.18	6.32	0.51	0.90
TR	5.90 ± 0.72	21.01 ± 3.34	16.86 ± 2.11	10.00 ± 1.02	9.97 ± 0.93	0.66 ± 0.10	1.60 ± 0.22
HCC	4.31 ± 0.46	23.35 ± 2.07	12.28 ± 0.96	7.71 ± 0.72	7.20 ± 0.60	0.53 ± 0.07	1.25 ± 0.11
LayeringTree	5.07 ± 0.25	25.48 ± 0.60	7.76 ± 0.54	2.97 ± 0.26	4.08 ± 0.27	–	–
Gromov	3.33 ± 0.45	13.28 ± 0.61	9.34 ± 0.53	4.08 ± 0.27	5.54 ± 0.49	0.43 ± 0.02	1.01 ± 0.04
DELTAZERO	**1.87** ± 0.08	**10.31** ± 0.62	**7.59** ± 0.38	**2.79** ± 0.15	**3.56** ± 0.20	**0.24** ± 0.00	**0.70** ± 0.03
Improvement (%)	43.8%	22.3%	2.3%	6.0%	12.7%	44.1 %	22.2%

✦ Sensitivity analysis