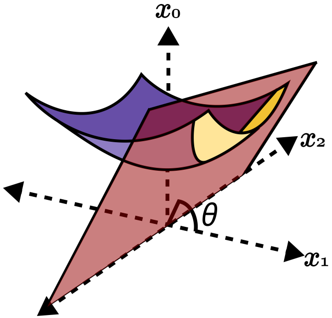
Fast hyperboloid decision tree algorithms

Philippe Chlenski,[1] Ethan Turok,[1] Antonio Moretti,[2] Itsik Pe'er[1]

[1]Columbia University [2]Barnard College



Motivation

Hyperbolic space: neighborhoods grows exponentially

Euclidean space cannot represent hierarchical data!

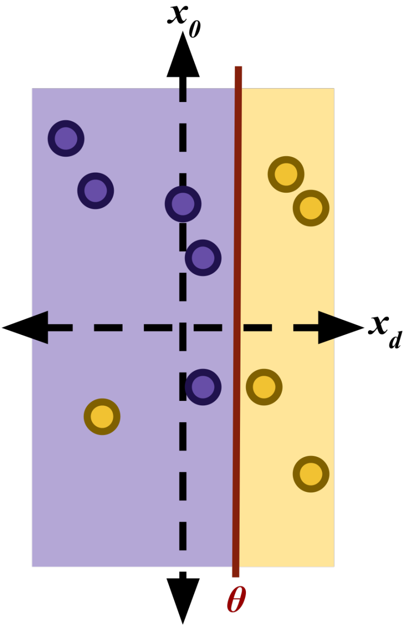
However, inference methods lag behind

Decision tree algorithms

Decision trees:

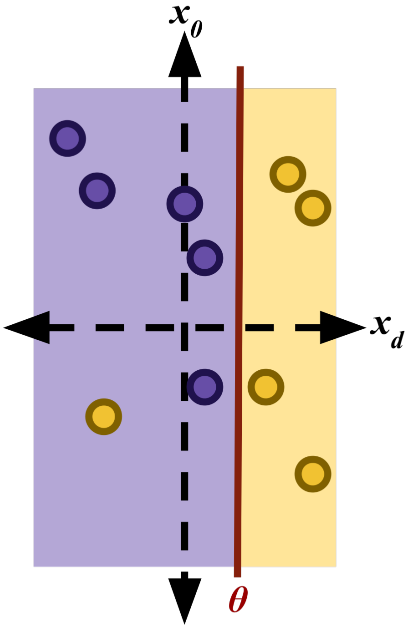
- Partition space recursively
- Split criteria: $x_d > \theta$?
- Decision areas are high-dimensional boxes
- Choose splits to maximize homogeneity

Random forests:
Ensemble of decision trees trained on subsets of data/features



Decision boundary wishlist

1. Topological continuity of decision areas
2. Convexity of decision areas
3. Equidistance to the points being separated
4. $O(nd)$ candidates per split

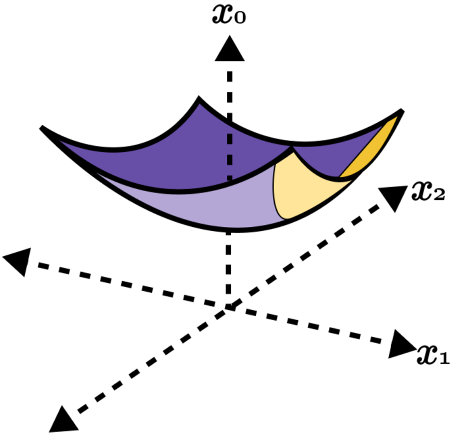


Problem setup

Inputs:

- **X**: points in hyperbolic space
- **y**: class labels

Task:
Fit a decision tree to predict **y** from **X**

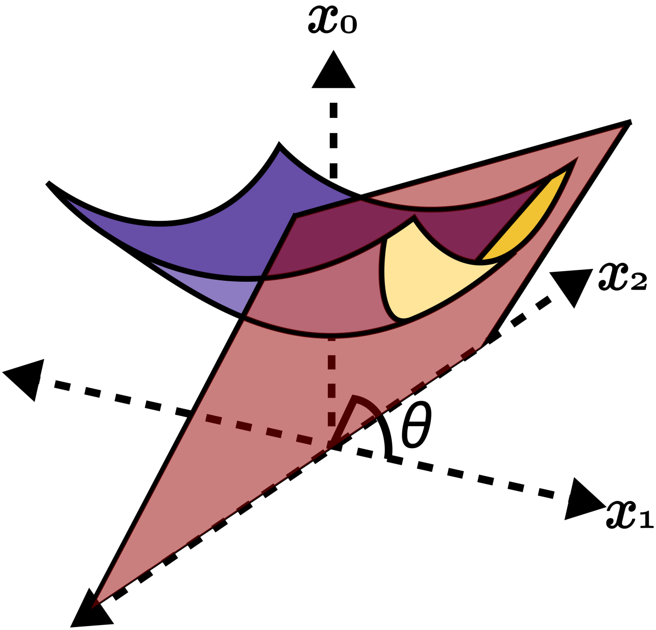


Decision trees interpretation

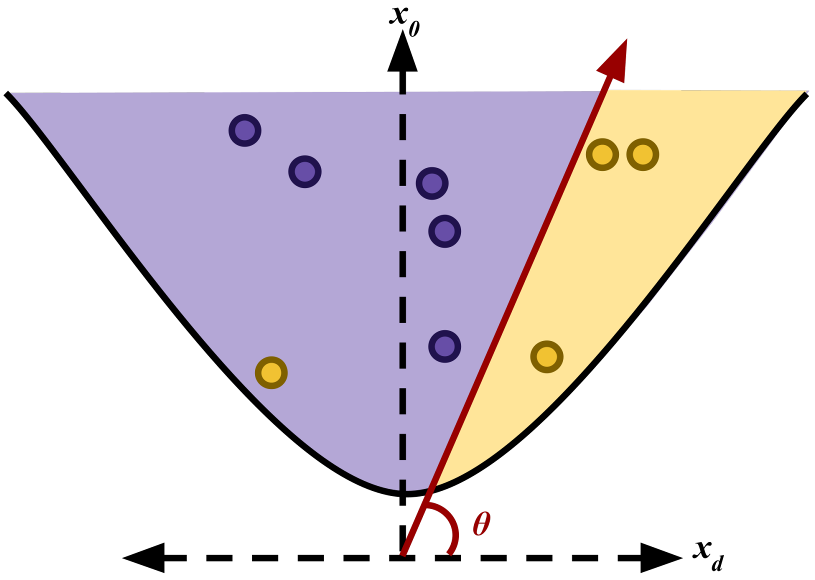
Decision tree boundaries = hyperplanes

Are there other hyperplanes we want to consider?

Do any of them fulfill the wishlist?



x_0
x_2
θ
x_1



x_0
θ
x_d

Implementation and extensions

HyperRF: ensemble of HyperDT decision trees (random forest)

`pip install hyperdt`

Classification and regression on the hyperboloid model using scikit-learn API

Source, experiments at
`https://github.com/pchlenski/hyperdt`

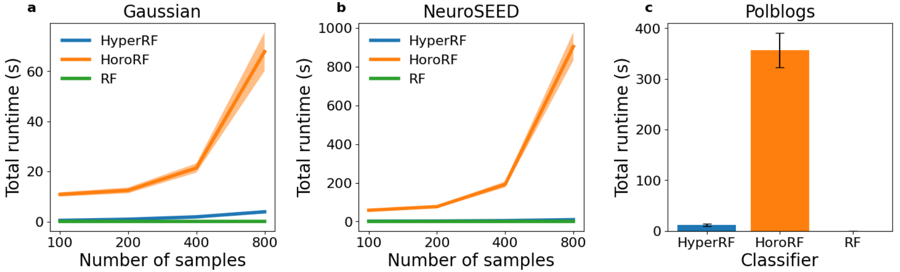
Gaussian results

		Decision Trees			Random Forests		
D	n	HYPERDT	Euclidean	HORODT	HYPERRF	Euclidean	HORORF
2	100	**89.10**†	87.90	84.60	**90.70**‡†	87.50	86.30

Gaussian results

		Decision Trees			Random Forests		
D	*n*	HYPERDT	Euclidean	HORODT	HYPERRF	Euclidean	HORORF
2	100	**89.10**†	87.90	84.60	**90.70**‡†	87.50	86.30
	200	**90.05**†	89.55	84.60	**90.60**	89.15	89.10
	400	**90.97**‡†	89.53	85.55	**91.32**‡†	89.00	88.88
	800	**91.88**‡†	90.14	85.75	**91.99**‡†	89.33	89.45
4	100	**98.70**†	97.70	93.60	**98.40**	97.90	97.90
	200	**98.75**‡†	98.10	95.80	**98.85**‡†	97.90	98.05
	400	**99.25**‡†	98.25	96.92	**99.30**‡†	98.22	98.50
	800	**99.30**‡†	98.36	97.27	**99.36**‡†	98.21	98.76
8	100	**99.70**†	99.60	97.70	**99.70**	99.50	99.10
	200	**99.65**†	99.60	98.20	**99.75**	99.70	**99.75**
	400	**99.90**†	99.88	99.10	99.88	**99.93**	99.88
	800	**99.96**†	99.90	99.38	**99.96**	99.91	99.94
16	100	**99.80**†	99.50	98.80	**99.80**	99.60	99.60
	200	99.95	**100.00**†	99.50	99.90	**99.95**	99.80
	400	**100.00**†	99.97	99.90	**100.00**	**100.00**	99.95
	800	**100.00**	99.99	99.90	**100.00**	99.99	99.92

Time complexity



Other results

Benchmarks on other datasets:

- Biological sequence embeddings
- Graph embeddings
- WordNet embeddings

Comparisons to other hyperbolic classifiers

Comparison to other models of hyperbolic space

Ablations

Conclusion

HYPERDT satisfies all decision boundary wishlist items

Sparse dot products maintain the asymptotic complexity of Euclidean decision trees

High accuracy in all inference settings

Easy to use

Future work

Extension to elliptical geometry and product space manifolds

Performance optimizations (e.g. pruning, caching)

Extension to more complex decision tree/random forest algorithm (e.g. boosted trees, branch-and-bound methods)

Hyperbolic data analysis with HYPERDT