Hyperbolic Embeddings of *Supervised Models*

Google Research
Google DeepMind

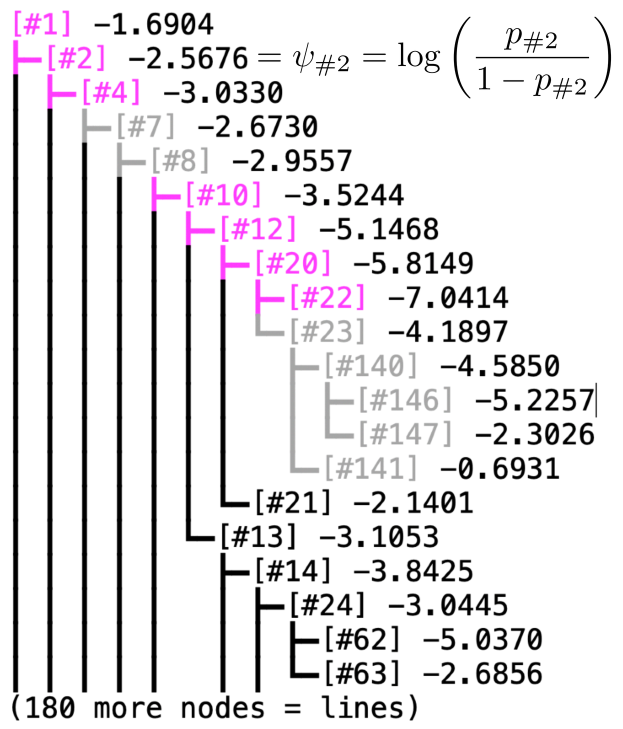
Richard Nock Ehsan Amid Frank Nielsen Alexander Soen Manfred K. Warmuth

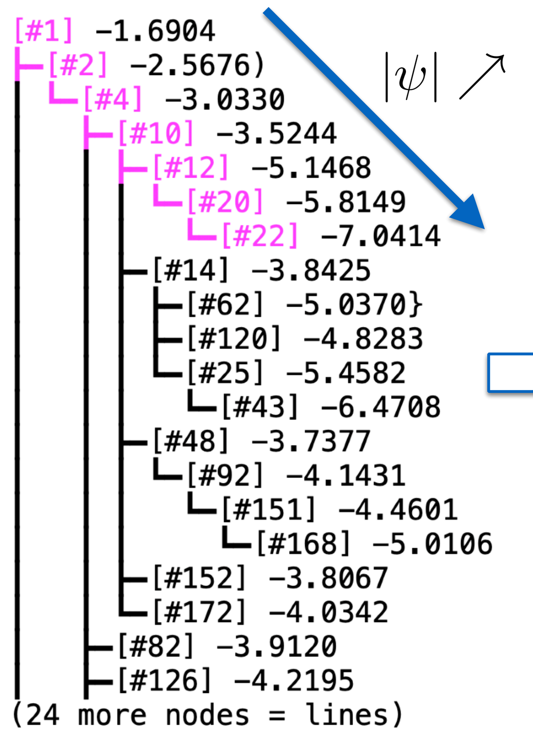
RIKEN AIP Australian National University Sony CSL

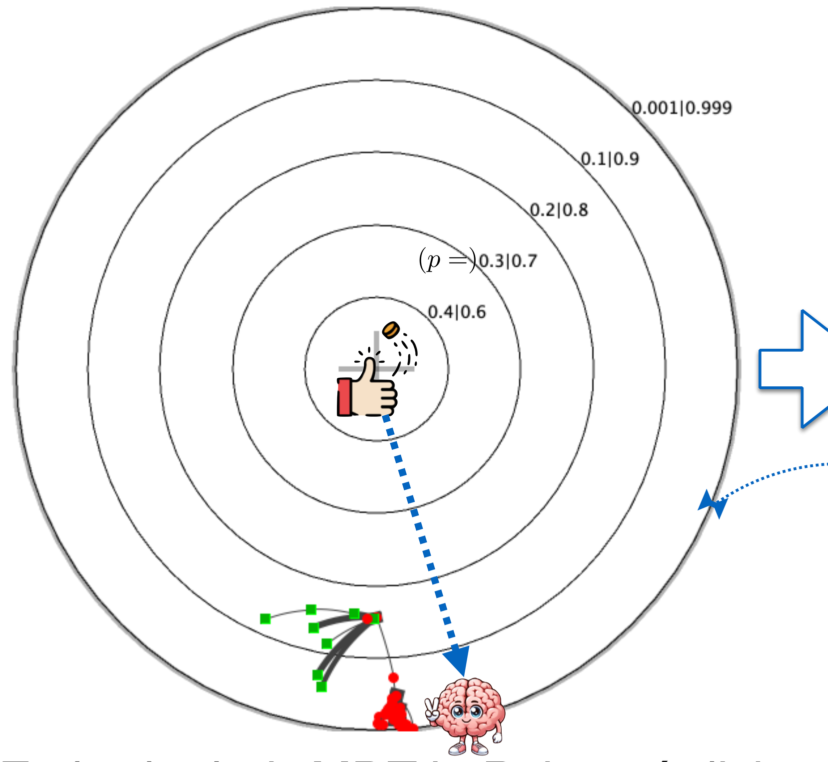
Summary

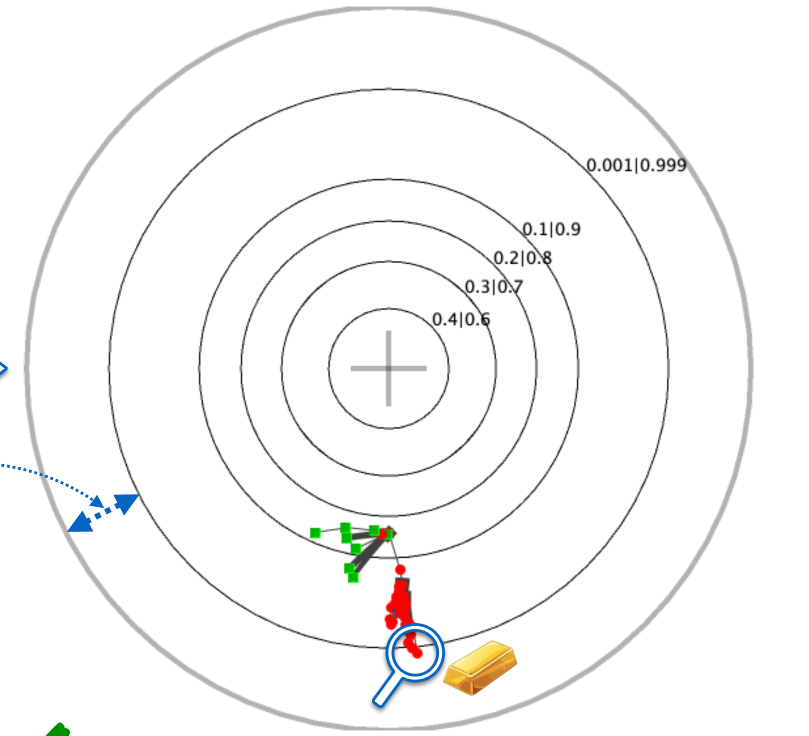
- Models of hyperbolic geometry models used in ML for two purposes:
  - embed **data**
  - embed **unsupervised** models (e.g. hierarchies)
- We demonstrate their capabilities to embed **supervised models**: (ensembles of) decision trees (DTs). To get there, we solve 3 problems:
  1. extract **monotonic** decision trees from DTs
  2. embed trees in Poincaré disk (d(origin,**node**) = absolute confidence of prediction for **log-loss**)
  3. smoothly bend distance → **readability**↗ & tackle known **numerical issue** of Poincaré disk
- Last step involves a generalization of Leibniz-Newton's fundamental theorem of calculus of independent interest

Decision Tree (DT) | ① Monotonic Decision Tree (MDT) | ② Embedding in Poincaré disk $\mathbb{B}$ | ③ t-self of Poincaré disk $\mathbb{B}^{(t)}$









- DT = hard to interpret / visualize:
  - ✗ Non monotonic rules (fct of confidences $|\psi_.|$)
  - ✗ No global visualization (BIG, many numbers)
- Extract monotonic subtree
  - ✓ monotonic rules ($|\psi| = \text{monotonic\_fct}(size)$)
  - ✗ No global visualization (big, many numbers)
- Embed *whole* MDT in Poincaré disk:
  - ✓ dist. to origin of node $= |\alpha_.|$ → faithful embed.
  - ✓ center: $p = 1/2$ = , border: $p \in \{0,1\}$ =
  - ✓ equidistant isolines
- Best part of model close to border
  - ✗ ↗↗ numerical inaccuracies, little readability
- ✓ Generalization of Leibniz-Newton's fundamental theorem of calculus
  - smoothly changes integral (→ distances)
  - ✓ tempered Poincaré disk $\mathbb{B}^{(t)}$ hyperbolic
  - ✓ stretches region close to border → readability ↗
  - ✓ ↘ non-linear stretch close to center

Contributions illustrated

Take home

3 key contributions, independent reusability:

- Link supervised learning and distances in hyperbolic geometry→embed more models!
- New monotonic decision tree models, extracted from decision trees→benefits+
- New fundamental theorem of t-calculus → general use in ML (integral ML distortions: Bregman & f-divergences, IPMs, OT, etc.)

More in paper:

- How to embed ensembles of boosted DTs
- Modified Sarkar's embedding for MDTs
- Application to Lorentz model
- Integration extended to *tempered*: properties & in-context analysis (geometry, ML):
  - additivity, dilativity, Chasles relationship, etc.
  - chain rule, mean value theorem, etc.
  - generalized hyperbolic Pythagorean theorem
  - data processing & convexity properties, etc.

Code & more: https://richardnock.github.io/

Class probability estimation vs Poincaré disk model

- Class Probability Estimation:
- 2 classes ($y = \pm 1$), model H → **posterior** $p = \mathbb{P}[y|\boldsymbol{x}, H]$
- **Log**-loss incurred (pointwise):

$$\underline{L}^{\text{LOG}}(p) = \begin{cases} -p \cdot \log p \\ -(1-p)\cdot\log(1-p) \end{cases}$$

- Duality with real-valued classification *via* the **canonical link** of the loss:

$$\psi_{\text{LOG}}(p) = \log\left(\frac{p}{1-p}\right)$$

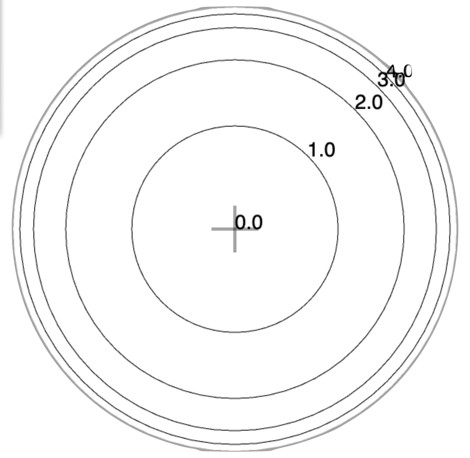
- Poincaré disk model $\mathbb{B}$:
- 2D hyperbolic geometry model
- Conformal, geodesics = circular arcs ($\perp$ to $\partial\mathbb{B}$ if prolonged)
- Hyperbolicity → low distortion embedding for **everything-tree**
- For any $\boldsymbol{z} \in \mathbb{B}$ with **Euclidean** norm $\|\boldsymbol{z}\| = r$, the distance to the center of the disk is

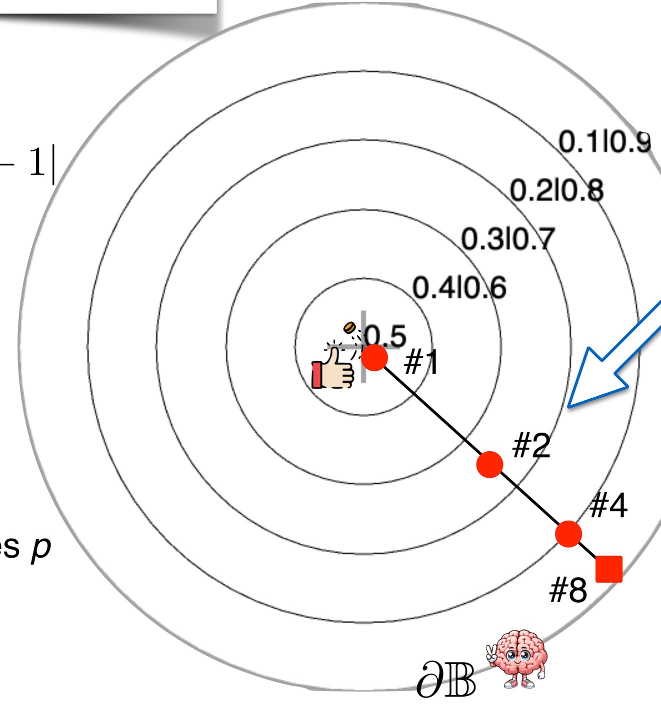
$$d_{\mathbb{B}}(\boldsymbol{z}, \boldsymbol{0}) = \log\left(\frac{1+r}{1-r}\right)$$
$$\|\boldsymbol{z}\| = r$$

- <u>Classification **confidence**</u>:

$$|\psi_{\text{LOG}}(p)| = \log\left(\frac{1+r}{1-r}\right)$$
$$r \doteq |2p-1|$$

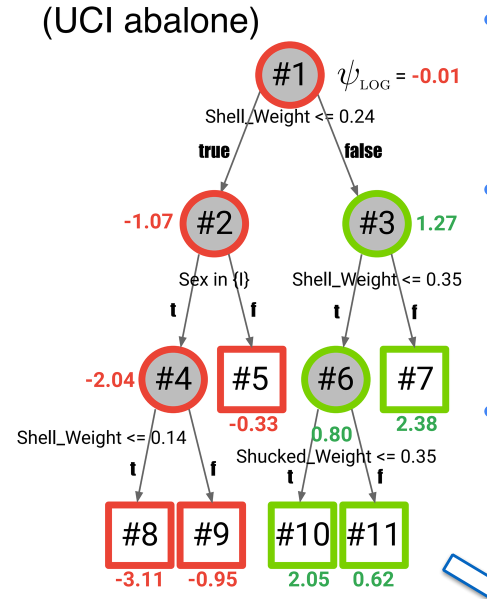
→ embed prediction p with some z such that $\|z\| = |2p-1|$



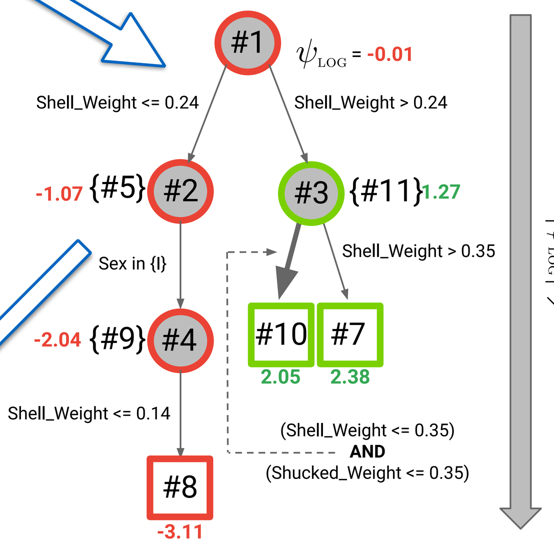


Monotonic Decision Trees

(UCI abalone)



- A DT = tree + posterior prediction at each node → good candidate model
- BUT $|\psi_{\text{LOG}}|$ **not** monotonic on paths → risks of unreadable / messy embeddings
- We use its **monotonic decision tree** (MDT) → *Cf* paper (algo., prop.)



Best part of model close to $\partial\mathbb{B}$ but this region is numerically error-prone (↗ risk of mapping to $\partial\mathbb{B}$)

Need low-level machinery!

Tempered calculus 101

- **Basic idea:** generalize Riemann summation in Riemann integration framework

Riemann integration framework

$z \oplus_t z' \doteq z + z' + (1-t)zz'$ ($t \in \mathbb{R}$, tempered addition)
- Nivanen *et al.*, 2003

$$\rightarrow {}^{(t)}\!\!\int_a^b f(x)\mathrm{d}_t x = t\text{-Riemann integration}$$

(t=1 → classical Riemann integration)

- **Theorem 1**: any f is t-Riemann integrable for *all* $t \in \mathbb{R}$ or for none. In the former case, we have

$$^{(t)}\!\!\int_a^b f(u)\mathrm{d}_t u = \mathfrak{g}_t\left(\int_a^b f(u)\mathrm{d}u\right) \text{ with } \mathfrak{g}_t(z) \doteq \log_t \exp z$$

$\log_t \doteq \frac{z^{1-t}-1}{1-t}$ ($t \neq 1$)
$\log_1 \doteq \log$

- **Theorem 2**: when it exists, let $\mathrm{D}_t f(z) \doteq \lim_{\delta \to 0} \left(f(z+\delta) \ominus_t f(z)\right)/\delta$.

tempered (t-) subtraction
$$z \ominus_t z' \doteq \frac{z - z'}{1 + (1-t)z'}$$

Suppose f t-Riemann integrable. Then, the function defined by $F(z) \doteq {}^{(t)}\!\!\int_a^z f(u)\mathrm{d}_t u$ is such that $\mathrm{D}_t F = f$.

F is a t-primitive of f (zeroes in z=a)
f is the t-derivative of F

Trivia: can you guess $\mathrm{D}_t \mathfrak{g}_t(z)$?

t-self of Poincaré disk

- From Theorem 1, distance becomes

$$d_{\mathbb{B}^{(t)}}(\boldsymbol{0}, \boldsymbol{z}) = \log_t \exp d_{\mathbb{B}}(\boldsymbol{0}, \boldsymbol{z}) = \log_t\left(\frac{1+\|\boldsymbol{z}\|}{1-\|\boldsymbol{z}\|}\right)$$

- Hyperbolicity & other properties: paper
- For $t \in [0,1]$, "pushes back $\partial\mathbb{B}$" & low non-linear distortion close to center: