Hyperbolic Embeddings for Preserving Privacy and Utility in Text

Oluwaseyi Feyisetan
Amazon
sey@amazon.com

Tom Diethe
Amazon
tdiethe@amazon.co.uk

Thomas Drake
Amazon
draket@amazon.com

ABSTRACT

Guaranteeing a certain level of user privacy in an arbitrary piece of text is a challenging issue. However, with this challenge comes the potential of unlocking access to vast data stores for training machine learning models and supporting data driven decisions. We address this problem through the lens of d_χ-privacy, a generalization of Differential Privacy to non Hamming distance metrics. In this work, we explore word representations in Hyperbolic space as a means of preserving privacy in text. We provide a proof satisfying d_χ-privacy, then we define a probability distribution in Hyperbolic space and describe a way to sample from it in high dimensions. Privacy is provided by perturbing vector representations of words in high dimensional Hyperbolic space to obtain a semantic generalization. We conduct a series of experiments to demonstrate the tradeoff between privacy and utility. Our privacy experiments illustrate protections against an authorship attribution algorithm while our utility experiments highlight the minimal impact of our perturbations on several downstream machine learning models. Compared to the Euclidean baseline, we observe > 20x greater guarantees on expected privacy against comparable worst case statistics.

ACM Reference Format:
Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2020. Hyperbolic Embeddings for Preserving, Privacy and Utility in Text. In *Proceedings of Workshop on Privacy and Natural Language Processing (PrivateNLP '20).* Houston, TX, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

PrivateNLP '20, February 7, 2020, Houston, TX, USA

Hyperbolic Metric Differential Privacy

Oluwaseyi Feyisetan, Tom Diethe, Thomas Drake
{sey,tdiethe,draket}@amazon.com

Summary

- **User's goal**: meet some specific need with respect to a query x
- **Agent's goal**: satisfy the user's request
- **Question**: what occurs when x is used to make other inferences
- **Mechanism**: Modify the query to protect privacy whilst preserving semantics
- **Our approach**: *Hyperbolic Metric Differential Privacy*

Metric Differential Privacy

A *randomised* mechanism $\mathcal{M} : X \to Y$ is (ε, δ)-differentially private if for all neighbouring inputs $x \simeq x'$ (i.e. $d_h(x, x') = 1$ where d_h is the Hamming distance) and for all sets of outputs $E \subseteq Y$, for $\delta \in [0, 1]$,

$$\mathbb{P}[\mathcal{M}(x) \in E] \leq e^{\varepsilon d_h(x,x')} \mathbb{P}\left[\mathcal{M}\left(x'\right) \in E\right] + \delta.$$

Metric DP generalizes this to use *any valid metric* $d(x, x')$, (i.e. satisfies nonnegativity, indiscernibles, symmetry, triangle inequality).
Mechanism: For Euclidean metric DP, we use multivariate Laplacian noise to achieve $\varepsilon-$mDP, i.e: $\boldsymbol{\xi} \sim \text{Lap}\left(\frac{1}{n\varepsilon}\right)$

- **Robust to post-processing**: $\mathcal{M}$ is (ε, δ)-DP, then $f(\mathcal{M})$ is at least (ε, δ)-DP
- **Composition**: if $\mathcal{M}_1, ..., \mathcal{M}_n$ are (ε, δ)-DP, $g(\mathcal{M}_1, ..., \mathcal{M}_n)$ is $(\sum_{i=1}^{n} \epsilon_i, \sum_{i=1}^{n} \delta_i)$-DP
- **Protects against side knowledge**: if attacker has prior $P^{x_i}_{prior}$ and computes $P^{x_i}_{posterior}$ after observing $\mathcal{M}(\mathbf{x})$ from ε-DP mechanism, then $dist(P^{x_i}_{prior}, P^{x_i}_{posterior}) = \mathcal{O}(\varepsilon)$.

Metric DP Mechanism for word embeddings

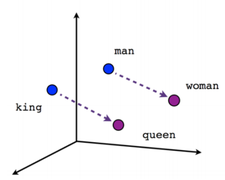
Inputs:
- $w \in \mathcal{W}$: word to be "privatised" from dictionary
- $\phi : \mathcal{W} \mapsto \mathcal{Z}$: embedding function
- $d : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$: distance function in embedding space
- $\Omega(\varepsilon)$: DP noise distribution

1. Project word $\mathbf{v} = \phi(\mathbf{w})$
2. Perturb the word vector: $\mathbf{v}' = \mathbf{v} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \Omega(\varepsilon)$
3. The new vector $\mathbf{v}'$ will not be a word (a.s.)
4. Project back to $\mathcal{W}$: $\mathbf{w}' = \arg\min_{\mathbf{w} \in \mathcal{W}} d(\mathbf{v}', \phi(\mathbf{w}))$
5. return $\mathbf{w}'$

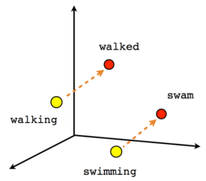
What do we need?
- d satisfies the axioms of a *metric*
- *A way to sample* using Ω in the metric space that respects d and gives us ε-metric DP

Hyperbolic Word Embeddings

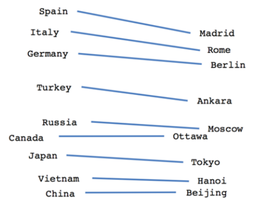
Traditional embeddings map from words into a vector space $\phi : \mathcal{W} \mapsto \mathbb{R}^n$, such as neural network based models (e.g. Word2Vec, GloVe, fastText). In this space, nearest neigbors preserve semantics.



Male-Female



Verb tense



Country-Capital

Hyperbolic space can be thought of as the continuous analog of a tree structure. In natural language, this captures *hypernomy* and *hyponomy*, leading to embeddings require fewer dimensions. Nearest neighbors are often hypernyms!

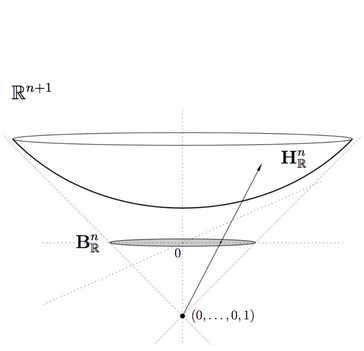


Fig. 1: Projection from Lorentz model $\mathcal{H}^n$ to Poincaré model

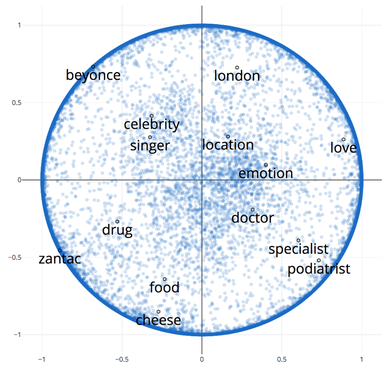


Fig. 2: WebIsADb IS-A relationships in GloVe on $\mathcal{B}^2$ Poincaré disk

- Distances in $n-$dimensional Poincaré ball model are given by:

$$d_{\mathcal{B}^n}(\mathbf{u}, \mathbf{v}) = \text{arcosh}\left(1 + 2\frac{\|\mathbf{u} - \mathbf{v}\|^2}{(1 - \|\mathbf{u}\|^2)(1 - \|\mathbf{v}\|^2)}\right)$$

- We prove that $d_{\mathcal{B}^n}(\mathbf{u}, \mathbf{v})$ is a valid metric (via Lorentzian model)

Hyperbolic Differential Privacy

- We derive the **Hyperbolic Laplace distribution** that satsifies ε-DP:

$$p(x|\mu = 0, \varepsilon) = \frac{1 + \varepsilon}{2\ {}_2F_1(1, \varepsilon, 2 + \varepsilon, -1)}\left(-\frac{2}{\|x\| - 1} - 1\right)^{-\varepsilon}$$

where ${}_2F_1(a, b; c, z)$ is the hypergeometric function
- We develop a Metropolis Hastings sampler that operates in Hyperbolic space

Experiments

To help choose ε, we define:
- Uncertainty statistics for the adversary over the outputs
- Indistinguishability statistics: plausible deniability
- Find a *radius of high protection*: guarantee on the likelihood of changing any word in the embedding vocabulary

Privacy Experiments 1
- **Task**: obfuscation vs. Koppel's authorship attribution algorithm
- **Datasets**: TPAN@Clef, correct author predictions (lower=better)

	PAN-11		PAN-12			
	small	large	set-A	set-C	set-D	set-I
0.5	36	72	4	3	2	5
1	35	73	3	3	2	5
2	40	78	4	3	2	5
8	65	116	4	5	4	5
∞	147	259	6	6	6	12

Privacy Experiments 2
- **Task**: expected privacy vs Euclidean baseline (lower N_w is better)
- **Datasets**: 100/200/300d GloVe embeddings

		expected value N_w			
ε	worst-case N_w	HYP-100	EUC-100	EUC-200	EUC-300
0.125	134	**1.25**	38.54	39.66	39.88
0.5	148	**1.62**	42.48	43.62	43.44
1	172	**2.07**	48.80	50.26	53.82
2	297	**3.92**	92.42	93.75	90.90
8	960	**140.67**	602.21	613.11	587.68

Utility Experiments
- **Tasks**: 5x classification (sentiment x2, product reviews, opinion polarity, question-type), 3x natural language tasks (NL inference, paraphrase detection, semantic textual similarity)
- **Baselines**: SentEval vs. random replacement

		HYP-100d			*original*		
Dataset	rand.	$\varepsilon = 1/8$	$\varepsilon = 1$	$\varepsilon = 8$	InferSent	SkipThought	fastText
MR	58.19	58.38	63.56	74.52	81.10	79.40	78.20
CR	77.48	83.21	83.92	85.19	86.30	83.1	80.20
MPQA	84.27	88.53	88.62	88.98	90.20	89.30	88.00
SST-5	30.81	41.76	42.40	42.53	46.30	-	45.10
TREC-6	75.20	82.40	82.40	84.20	88.20	88.40	83.40
SICK-E	79.20	81.00	82.38	82.34	86.10	79.5	78.9
MRPC	69.86	74.78	75.07	75.01	76.20	-	74.40
STS14	0.17	0.44	0.45	0.52	0.68	0.44	0.65

Selected References

[1] K.Chatzikokolakis, et al. Broadening the scope of differential privacy using metrics. Intl. Symposium on Privacy Enhancing Technologies, 2013.
[2] M. Nickel and D. Kiela, Poincaré embeddings for learning hierarchical representations. NeurIPS, 2017.