Hyperbolic Knowledge Graph Embeddings for Knowledge Base Completion

Prodromos Kolyvakis[1], Alexandros Kalousis[2], and Dimitris Kiritsis[1]

[1] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{prodromos.kolyvakis,dimitris.kiritsis}@epfl.ch
[2] Business Informatics Department, University of Applied Sciences, Western Switzerland Carouge, HES-SO, Switzerland
alexandros.kalousis@hesge.ch

Abstract. Learning embeddings of entities and relations existing in knowledge bases allows the discovery of hidden patterns in them. In this work, we examine the contribution of geometrical space to the task of knowledge base completion. We focus on the family of translational models, whose performance has been lagging. We extend these models to the hyperbolic space so as to better reflect the topological properties of knowledge bases. We investigate the type of regularities that our model, dubbed *HyperKG*, can capture and show that it is a prominent candidate for effectively representing a subset of Datalog rules. We empirically show, using a variety of link prediction datasets, that hyperbolic space allows to narrow down significantly the performance gap between translational and bilinear models and effectively represent certain types of rules.

Keywords: Knowledge Graph Embeddings · Hyperbolic Embeddings · Knowledge Base Completion

1 Introduction

Learning in the presence of structured information is an important challenge for artificial intelligence [18, 31, 41]. Knowledge Bases (KBs) such as WordNet [29], Freebase [8], YAGO [47] and DBpedia [27] constitute valuable such resources needed for a plethora of practical applications, including question answering and information extraction. However, despite their formidable number of facts, it is widely accepted that their coverage is still far from being complete [44, 58]. This shortcoming has opened the door for a number of studies addressing the problem of automatic knowledge base completion (KBC) or link prediction [34]. The impetus of these studies arises from the hypothesis that statistical regularities lay in KB facts, which when correctly exploited can result in the discovery of missing true facts [60]. Building on the great generalisation capability of distributed representations, a great line of research [10, 35, 36, 51, 62] has focused on learning KB vector space embeddings as a way of predicting the plausibility of a fact.

An intrinsic characteristic of knowledge graphs is that they present power-law (or scale-free) degree distributions as many other networks [15, 46]. In an attempt of understanding scale-free networks' properties, various generative models have been proposed such as the models of Barabási and Albert [6] and Van Der Hofstad [53]. Interestingly, Krioukov et al. [25] have shown that scale-free networks naturally emerge in the hyperbolic space. Recently, the hyperbolic geometry was exploited in various works [17, 37, 38, 42] as a means to provide high-quality embeddings for hierarchical structures. Hyperbolic space has the potential to bring significant value in the task of KBC since it offers a natural way to take the KB's topological information into account. Furthermore, many of the relations appearing in KBs lead to hierarchical and hierarchical-like structures [28].

At the same time, the expressiveness of various KB embedding models has been recently examined in terms of their ability to express any ground truth of facts [23, 56]. Moreover, Gutiérrez-Basulto and Schockaert [21] have proceeded one step further and investigated the compatibility between ontological axioms and different types of KB embeddings. Specifically, the authors have proved that a certain family of rules, i.e., the quasi-chained rules which form a subset of Datalog rules [1], can be exactly represented by a KB embedding model whose relations are modelled as convex regions; ensuring, thus, logical consistency in the facts induced by this KB embedding model. In the light of this result, it seems important that the appropriateness of a KB embedding model should not only be measured in terms of fully expressiveness but also in terms of the rules that it can model.

In this paper, we explore geometrical spaces having the potential to better represent KBs' topological properties and rules and examine the performance implications on KBC. We focus on the family of translational models [10] that attempt to model the statistical regularities as vector translations between entities' vector representations, and whose performance has been lagging. We extend the translational models by learning embeddings of KB entities and relations in the Poincaré-ball model of hyperbolic geometry. We do so by learning compositional vector representations [30] of the entities appearing in a given fact based on translations. The implausibility of a fact is measured in terms of the hyperbolic distance between the compositional vector representations of its entities and the learned relation vector. We prove that the relation regions captured by our proposed model are convex. Our model becomes, thus, a prominent candidate for representing effectively quasi-chained rules.

Among our contributions is the proposal of a novel KB embedding model as well as a regularisation scheme on the Poincaré-ball model, whose effectiveness we prove empirically. Furthermore, we prove that translational models do not suffer from the restrictions identified by Kazemi and Poole [23] in the case where a fact is considered valid when its implausibility score is below a certain non-zero threshold. We evaluate our approach on various benchmark datasets and our experimental results show that our work makes a big step towards (i) closing the performance gap between translational and bilinear models and (ii) enhancing our understanding of which KBs mostly benefit from exploiting hyperbolic embeddings. Last but not least, our work demonstrates that the choice of geometrical space plays a significant role for KBC and illustrates the importance of taking both the topological and the formal properties of KBs into account. The implementation code and the datasets are publicly available on: `https://github.com/prokolyvakis/hyperkg`.

2 Related Work

Shallow KB Embedding Models. There has been a great line of research dedicated to the task of learning distributed representations for entities and relations in KBs. To constrain the analysis, we only consider shallow embedding models that do not exploit deep neural networks or incorporate additional external information beyond the KB facts. For an elaborated review of these techniques, please refer to Nickel et al. [34] and Wang et al. [55]. We also exclude from our comparison recent work that explores different types of training regimes such as adversarial training, and/or the inclusion of reciprocal facts [11, 23, 26, 48] to make the analysis less biased to factors that could overshadow the importance of the geometrical space.

In general, the shallow embedding approaches can be divided into two main categories; the translational [10] and the bilinear [36] family of models. In the translational family, the vast majority of models [13, 22, 57, 59] generalise TransE [10], which attempts to model relations as translation operations between the vector representations of the *subject* and *object* entities, as observed in a given fact. In the bilinear family, most of the approaches [35, 51, 62] generalise RESCAL [36] that proposes to model facts through bilinear operations over entity and relations vector representations. In this paper, we focus on the family of translational models, whose performance has been lagging, and propose extensions in the hyperbolic space which by exploiting the topological and the formal properties of KBs bring significant performance improvements.

Hyperbolic Embeddings. There has been a growing interest in embedding scale-free networks in the hyperbolic space [7, 39]. Hyperbolic geometry was also exploited in various works as a way to exploit hierarchical information and learn more efficient representations [17, 37, 38, 42]. However, this line of work has only focused on single-relational networks. Recently and in parallel to our work, two other works have explored hyperbolic embeddings for KBs. Contrary to our work where Möbius or Euclidean addition is used as a translational operation, Suzuki et al. [49] exploit vector fields with an attractive point to generalise translation in Riemannian manifolds. Their approach, although promising, shows a degraded performance on commonly used benchmarks. Similarly to our approach, Balažević et al. [5] extend to the hyperbolic space the family of translational models demonstrating significant performance improvements over state-of-the-art. However, the authors exploit both the hyperbolic as well as the Euclidean space by using the *Möbius Matrix-vector multiplication* and Euclidean scalar biases.[3] Unlike our experimental setup, the authors also include reciprocal facts.

[3] The matrix, used in Möbius multiplication, and the biases are defined on Euclidean space and are learned through Euclidean SGD.

Although their approach is beneficial for KBC, it becomes hard to quantify the contributions of hyperbolic space. This is verified by the fact that their Euclidean model analogue performs in line with their "hybrid" hyperbolic-Euclidean model. Finally, neither of these works studies the types of rules that their proposed models can effectively represent.

3 Methods

3.1 Preliminaries

We introduce some definitions and additional notation that we will use throughout the paper. We denote the vector concatenation operation by the symbol $\oplus$ and the inner product by $\langle \cdot, \cdot \rangle$. We define the *rectifier* activation function as: $[\cdot]_+ := \max(\cdot, 0)$.

Quasi-chained Rules. Let $\mathbf{E}, \mathbf{N}$ and $\mathbf{V}$ be disjoint sets of *entities, (labelled) nulls* and *variables*, respectively.[4] Let $\mathbf{R}$ be the set of relation symbols. A *term* t is an element in $\mathbf{E} \cup \mathbf{N} \cup \mathbf{V}$; an *atom* α is an expression of the form $R(t_1, t_2)$, where R is a *relation* between the terms t_1, t_2. Let $\mathsf{terms}(\alpha) := \{t_1, t_2\}$; $\mathsf{vars}(\alpha) := \mathsf{terms}(\alpha) \cap \mathbf{V}$ and B_n for $n \geq 0$, H_k for $k \geq 1$ be atoms with terms in $\mathbf{E} \cup \mathbf{V}$. Additionally, let $X_j \in \mathbf{V}$ for $j \geq 1$. A *quasi-chained (QC) rule* σ [21] is an expression of the form:

$$B_1 \wedge \ldots \wedge B_n \to \exists X_1, \ldots, X_j . H_1 \wedge \ldots \wedge H_k, \tag{1}$$

where for all $i : 1 \leq i \leq n$

$$|(\mathsf{vars}(B_1) \cup ... \cup \mathsf{vars}(B_{i-1})) \cap \mathsf{vars}(B_i)| \leq 1$$

The QC rules constitute a subset of Datalog rules. A *database* D is a finite set of *facts*, i.e., a set of atoms with terms in $\mathbf{E}$. A *knowledge base (KB)* $\mathcal{K}$ consists of a pair (Σ, D) where Σ is an ontology whose axioms are QC rules and D a database. It should be noted that no constraint is imposed on the number of available axioms in the ontology. The ontology could be minimal in the sense of only defining the relation symbols. However, any type of rule, whether it is the product of the ontological design or results from formalising a statistical regularity, should belong to the family of QC rules. The Gene Ontology [4] constitutes one notable example of an ontology that exhibits QC rules.

Circular Permutation Matrices. An orthogonal matrix is defined as a real square matrix whose columns and rows are orthogonal unit vectors (i.e., orthonormal vectors), i.e.,

$$Q^{\mathrm{T}} Q = Q Q^{\mathrm{T}} = I \tag{2}$$

where I is the identity matrix. Orthogonal matrices preserve the vector inner product and, thus, they also preserve the Euclidean norms. Let $1 \leq i < n$, we

[4] Only existential variables can be mapped to labelled nulls.

define the *circular permutation matrix* Π_i to be the orthogonal $n \times n$ matrix that is associated with the following circular permutation of a n-dimensional vector $\boldsymbol{x}$:

$$\begin{pmatrix} x_1 & \cdots & x_{n-i} & x_{n-i+1} & \cdots & x_n \\ x_{i+1} & \cdots & x_n & x_1 & \cdots & x_i \end{pmatrix} \tag{3}$$

where x_i is the ith coordinate of $\boldsymbol{x}$ and i controls the number of $n - i$ successive circular shifts.

Hyperbolic Space. In this work, we exploit the Poincaré-ball model of the hyperbolic geometry. The Poincaré-ball model is the Riemannian manifold $\mathbb{P}^n = (\mathbb{B}^n, d_p)$, where $\mathbb{B}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < 1\}$ and d_p is the distance function:

$$\begin{aligned} d_p(\boldsymbol{u}, \boldsymbol{v}) &= \operatorname{acosh}\left(1 + 2\delta(\boldsymbol{u}, \boldsymbol{v})\right) \\ \delta(\boldsymbol{u}, \boldsymbol{v}) &= \frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)} \end{aligned} \tag{4}$$

The Poincaré-ball model presents a group-like structure when it is equipped with the *Möbius addition* [40, 52], defined by:

$$\boldsymbol{u} \boxplus \boldsymbol{v} := \frac{(1 + 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle + \|\boldsymbol{v}\|^2)\boldsymbol{u} + (1 - \|\boldsymbol{u}\|^2)\boldsymbol{v}}{1 + 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle + \|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|^2} \tag{5}$$

The isometries of $(\mathbb{B}^n, d_p)$ can be expressed as a composition of a left gyrotranslation with an orthogonal transformation restricted to $\mathbb{B}^n$, where the *left gyrotranslation* is defined as $L_u : v \mapsto u \boxplus v$ [2, 40]. Therefore, circular permutations constitute zero-left gyrotranslation isometries of the Poincaré-ball model.

3.2 HyperKG

The database of a KB consists of a set of facts in the form of $R(subject, object)$. We will learn hyperbolic embeddings of entities and relations such that valid facts will have a lower implausibility score than the invalid ones. To learn such representations, we extend the work of Bordes et al. [10] by defining a translation-based model in the hyperbolic space; embedding, thus, both entities and relations in the same space.

Let $\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{o} \in \mathbb{B}^n$ be the hyperbolic embeddings of the *subject, relation and object*, respectively, appearing in the $R(subject, object)$ fact. We define a *term embedding* as a function $\xi \colon \mathbb{B}^n \times \mathbb{B}^n \to \mathbb{B}^n$, that creates a composite vector representation for the pair $(subject, object)$. Since our motivation is to generalise the translation models to the hyperbolic space, a natural way to define the term embeddings is by using the Möbius addition. However, we found out empirically that the normal addition in the Euclidean space generalises better than the Möbius addition. We provide a possible explanation for this behaviour in an ablation study presented in the Results & Analysis section. To introduce non-commutativity in the term composition function, we use a circular permutation matrix to project the object embeddings. Non-commutativity is important because it allows to

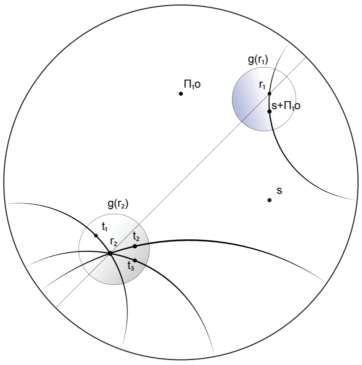

Fig. 1: A visualisation of HyperKG model in the $\mathbb{P}^2$ space. The geodesics of the disk model are circles perpendicular to its boundary. The zero-curvature geodesic passing from the origin corresponds to the line $\epsilon : y - x = 0$ in the Euclidean plane. Reflections over the line ϵ are equivalent to Π_1 permutations in the plane. $s, \Pi_1 o, s + \Pi_1 o$ are the subject vector, the permuted object vector and the composite term vector, respectively. $g(r_1), g(r_2)$ denote the geometric loci of term vectors satisfying relations R_1, R_2, with relation vectors $r1, r2$. t_1, t_2, t_3 are valid term vectors for the relation R_2.

model asymmetric relations with compositional representations [35]. Therefore, we define the term embedding as: $\boldsymbol{s} + \Pi_\beta \boldsymbol{o}$, where β is a hyperparameter controlling the number of successive circular shifts. To enforce the term embeddings to stay in the Poincaré-ball, we constrain all the entity embeddings to have a Euclidean norm less than 0.5. Namely, $\|\boldsymbol{e}\| < 0.5$ and $\|\boldsymbol{r}\| < 1.0$ for all entity and relation vectors, respectively. It should be noted that the entities' norm constraints do not restrict term embeddings to span the Poincaré-ball. We define the implausibility score as the hyperbolic distance between the term and the relation embeddings. Specifically, the implausibility score of a fact is defined as:

$$f_R(s, o) = d_p(\boldsymbol{s} + \Pi_\beta \boldsymbol{o}, \boldsymbol{r}) \tag{6}$$

Figure 1 provides an illustration of the HyperKG model in $\mathbb{P}^2$. We follow previous work [10] to minimise the following hinge loss function:

$$\mathcal{L} = \sum_{\substack{R(s,o)\sim P,\\ R'(s',o')\sim N}} [\gamma + f_R(s, o) - f_{R'}(s', o')]_+ \tag{7}$$

where P is the training set consisting of valid facts, N is a set of corrupted facts. To create the corrupted facts, we experimented with two strategies. We replaced randomly either the subject or the object of a valid fact with a random entity (but not both at the same time). We denote with $\#_{negs_E}$ the number of negative examples. Furthermore, we experimented with replacing randomly the relation while retaining intact the entities of a valid fact. We denote with $\#_{negs_R}$ the number of "relation-corrupted" negative examples. We employ the "*Bernoulli*" sampling method to generate incorrect facts [22, 57, 60].

As pointed out in different studies [10, 12, 26], regularisation techniques are really beneficial for the task of KBC. Nonetheless, very few of the classical regularisation methods are directly applicable or easily generalisable in the Poincaré-ball model of hyperbolic space. For instance, the ℓ_2 regularisation constraint imposes vectors to stay close to the origin, which can lead to underflows. The same holds for dropout [45], when a rather large dropout rate was used.[5] In our experiments, we noticed a tendency of the word vectors to stay close to the origin. Imposing a constraint to the vectors to stay away from the origin stabilised the training procedure and increased the model's generalisation capability. It should be noted that as the points in the Poincaré-ball approach the ball's boundary their distance $d_p(\boldsymbol{u}, \boldsymbol{v})$ approaches $d_p(\boldsymbol{u}, \boldsymbol{0}) + d_p(\boldsymbol{0}, \boldsymbol{v})$, which is analogous to the fact that in a tree the shortest path between two siblings is the path through their parent [42]. Building on this observation, our regulariser further imposes this "tree-like" property. Additionally, since the volume in hyperbolic space grows exponentially, our regulariser implicitly penalises crowding. Let $\Theta := \{\boldsymbol{e}_i\}_{i=1}^{|\mathbf{E}|} \bigcup \{\boldsymbol{r}_i\}_{i=1}^{|\mathbf{R}|}$ be the set of all entity and relation vectors, where $|\mathbf{E}|, |\mathbf{R}|$ denote the cardinalities of the sets $\mathbf{E}, \mathbf{R}$, respectively. $\mathcal{R}(\Theta)$ defines our proposed regularisation loss function:

$$\mathcal{R}(\Theta) = \sum_{i=1}^{|\mathbf{E}|+|\mathbf{R}|} (1 - \| \boldsymbol{\theta}_i \|^2) \tag{8}$$

The overall embedding loss is now defined as $\mathcal{L}'(\Theta) = \mathcal{L}(\Theta) + \lambda \mathcal{R}(\Theta)$, where λ is a hyperparameter controlling the regularisation effect. We define $a_i := 0.5$, if θ_i corresponds to an entity vector and $a_i := 1.0$, otherwise. To minimise $\mathcal{L}'(\Theta)$, we solve the following optimisation problem:

$$\Theta' \leftarrow \arg\min_{\Theta} \mathcal{L}'(\Theta) \qquad \text{s.t. } \forall\, \boldsymbol{\theta}_i \in \Theta : \|\boldsymbol{\theta}_i\| < a_i. \tag{9}$$

To solve Equation (9), we follow Nickel and Kiela [37] and use Riemannian SGD (RSGD; 9). In RSGD, the parameter updates are of the form:

$$\boldsymbol{\theta}_{t+1} = \mathfrak{R}_{\boldsymbol{\theta}_t} \left(-\eta \nabla_R \mathcal{L}'(\boldsymbol{\theta}_t) \right)$$

where $\mathfrak{R}_{\boldsymbol{\theta}_t}$ denotes the retraction onto the open d-dimensional unit ball at $\boldsymbol{\theta}_t$ and η denotes the learning rate. The Riemannian gradient of $\mathcal{L}'(\boldsymbol{\theta})$ is denoted by $\nabla_R \in \mathcal{T}_\theta \mathbb{B}$. The Riemannian gradient can be computed as $\nabla_R = \frac{(1-\|\boldsymbol{\theta}_t\|^2)^2}{4} \nabla_E$, where ∇_E denotes the Euclidean gradient of $\mathcal{L}'(\boldsymbol{\theta})$. Similarly to Nickel and Kiela [37], we use the following retraction operation $\mathfrak{R}_{\boldsymbol{\theta}}(\boldsymbol{v}) = \boldsymbol{\theta} + \boldsymbol{v}$.

To constrain the embeddings to remain within the Poincaré ball and respect the additional constraints, we use the following projection:

$$\text{proj}(\boldsymbol{\theta}, a) = \begin{cases} a\boldsymbol{\theta}/(\|\boldsymbol{\theta}\| + \varepsilon) & \text{if } \|\boldsymbol{\theta}\| \geq a \\ \boldsymbol{\theta} & \text{otherwise} \end{cases}, \tag{10}$$

[5] In our experiments, we noticed that a rather small dropout rate had no effect on the model's generalisation capability.

where ε is a small constant to ensure numerical stability. In all experiments we used $\varepsilon = 10^{-5}$. Let a be the constraint imposed on vector $\boldsymbol{\theta}$, the full update for a single embedding is then of the form:

$$\boldsymbol{\theta}_{t+1} \leftarrow \text{proj}\left(\boldsymbol{\theta}_t - \eta\frac{(1 - \|\boldsymbol{\theta}_t\|^2)^2}{4}\nabla_E, a\right). \tag{11}$$

We initialise the embeddings using the Xavier initialization scheme [19], where we use Equation (10) for projecting the vectors whose norms violate the imposed constraints. Finally, it should be noted that the space complexity of HyperKG is the same as that of TransE and, based on our measurements, the running time of HyperKG is almost double compared to that of TransE [10] and ComplEx [51].

3.3 Convex Relation Spaces

In this section, we investigate the type of rules that HyperKG can model. Recently, Wang et al. [56] proved that the bilinear models are universal, i.e., they can represent every possible fact given that the dimensionality of the vectors is sufficient. The authors have also shown that the TransE model is not universal. In parallel, Kazemi and Poole [23] have shown that the FTransE model [16], which is the most general translational model proposed in the literature, imposes some severe restrictions on the types of relations the translational models can represent. In the core of their proof lies the assumption that the implausibility score defined by the FTransE model approaches zero for all given valid facts. Nonetheless, this condition is less likely to be met from an optimisation perspective [59].

Additionally, Gutiérrez-Basulto and Schockaert [21] studied the types of regularities that KB embedding methods can capture. To allow for a formal characterisation, the authors considered hard thresholds λ_R such that a fact $R(s, o)$ is considered valid iff $s_R(\boldsymbol{s}, \boldsymbol{o}) \leq \lambda_R$, where $s_R(.,.)$ is the implausibility score. It should be highlighted that KB embeddings are often learned based on a maximum-margin loss function, which ideally leads to hard-threshold separation. The vector space representation of a given relation R can then be viewed as a region $n(R)$ in $\mathbb{R}^{2n}$, defined as follows:

$$n(R) = \{\boldsymbol{s} \oplus \boldsymbol{o} \,|\, s_R(\boldsymbol{s}, \boldsymbol{o}) \leq \lambda_R\} \tag{12}$$

Based on this view of the relation space, the authors prove that although bilinear models are fully expressive, they impose constraints on the type of rules they can learn. Specifically, let $R_1(X, Y) \rightarrow S(X, Y)$, $R_2(X, Y) \rightarrow S(X, Y)$ be two valid rules. The bilinear models impose either that $R_1(X, Y) \rightarrow R_2(X, Y)$ or $R_2(X, Y) \rightarrow R_1(X, Y)$; introducing, thus, a number of restrictions on the type of subsumption hierarchies they can model. Gutiérrez-Basulto and Schockaert [21], additionally, prove that there exists a KB embedding model with convex relation regions that can correctly represent knowledge bases whose axioms belong to the family of QC rules. Equivalently, any inductive reasoning made by the aforementioned KB embedding model would be logically consistent and deductively closed with respect to the ontological rules. It can be easily verified that the relation regions of TransE [10] are indeed convex. This result is in accordance with the results of Wang et al. [56]; TransE is not fully expressive. However, it could be a prominent candidate for representing QC rules consistently. Nonetheless, this result seems to be in conflict with the results of Kazemi and Poole [23]. Let $s_R^{TE}(s,o)$ be the implausibility score of TransE, we demystify this seeming inconsistency by proving the following lemma:

Lemma 1. *The restrictions proved by Kazemi and Poole [23] do not apply to the TransE model when a fact is considered valid iff* $s_R^{TE}(s,o) \leq \lambda_R$ *for sufficient* $\lambda_R > 0$.

We prove Lemma 1 in the Supplemental Material, which is also provided in [24], by constructing counterexamples for each one of the restrictions. Since the restrictions can be lifted for the TransE model, we can safely conclude that they are not, in general, valid for all its generalisations. In parallel, we built upon the formal characterisation of relations regions, defined in Equation (12) and we prove that the relation regions captured by HyperKG are indeed convex. Specifically, we prove:

Proposition 1. *The geometric locus of the term vectors, in the form of* $\boldsymbol{s} + \Pi_\beta \boldsymbol{o}$, *that satisfy the equation* $d_p(\boldsymbol{s} + \Pi_\beta \boldsymbol{o}, \boldsymbol{r}) \leq \lambda_R$ *for some* $\lambda_R > 0$ *corresponds to a d-dimensional closed ball in the Euclidean space. Let* $\rho = \frac{\cosh(\lambda_R)-1}{2}(1 - \|\boldsymbol{r}\|^2)$, *the geometric locus can be written as* $\|\boldsymbol{s} + \Pi_\beta \boldsymbol{o} - \frac{\boldsymbol{r}}{\rho+1}\|^2 \leq \frac{\rho}{\rho+1} + \frac{\|\boldsymbol{r}\|^2}{(\rho+1)^2} - \frac{\|\boldsymbol{r}\|^2}{\rho+1}$, *where the ball's radius is guaranteed to be strictly greater than zero.*

The proof of Proposition 1 can also be found in the Supplemental Material – also provided in [24]. By exploiting the triangle inequality, we can easily verify that the relation regions captured by HyperKG are indeed convex. Figure 1 provides an illustration of the geometric loci captured by HyperKG in $\mathbb{B}^2$. This result shows that HyperKG constitutes another one prominent embedding model for effectively representing QC rules.

4 Experiments

We evaluate our HyperKG model on the task of KBC using two sets of experiments. We conduct experiments on the WN18RR [12] and FB15k-237 [50] datasets. We also construct two datasets whose statistical regularities can be expressed as QC rules to test our model's performance in their presence. WN18RR and FB15k-237 constitute refined subsets of WN18 and FB15K that were introduced by Bordes et al. [10]. Toutanova and Chen [50] identified that WN18 and FB15K contained a lot of reversible relations, enabling, thus, various KB embedding models to generalise easily. Exploiting this fact, Dettmers et al. [12] obtained state-of-the-art results only by using a simple reversal rule. WN18RR and FB15k-237 were carefully created to alleviate this leakage of information.

To test whether the scale-free distribution provides a reasonable means for modelling topological properties of knowledge graphs, we investigate the degree

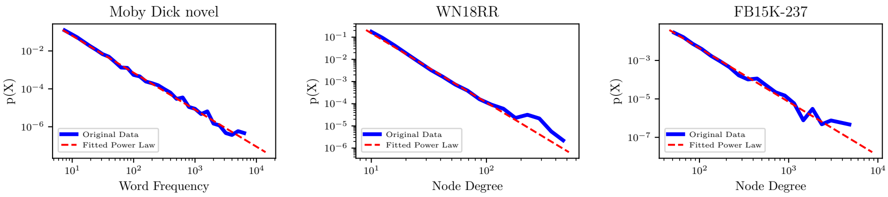

Fig. 2: A visualisation of the probability density functions using a histogram with log-log axes.

distributions of WN18RR and FB15k-237. Similarly to Steyvers and Tenenbaum [46], we treat the knowledge graphs as undirected networks. We also compare against the distribution of the frequency of word usage in the English language; a phenomenon that is known to follow a power-law distribution [63]. To do so, we used the frequency of word usage in Herman Melville's novel "Moby Dick" [32]. We followed the procedure described by Alstott et al. [3]. In Figure 2, we show our analysis where we demonstrate on a histogram with log-log axes the probability density function with regard to the observed property for each dataset, including the fitted power-law distribution. It can be seen that the power-law distribution provides a reasonable means for also describing the degree distribution of KBs; justifying the work of Steyvers and Tenenbaum [46]. The fluctuations in the cases of WN18RR and FB15k-237 could be explained by the fact that the datasets are subsets of more complete KBs; a fact that introduces noise which in turn can explain deviations from the perfection of a theoretical distribution [3].

4.1 Datasets

To test our model's performance on capturing QC rules, we extract from Wikidata [14, 54] two subsets of facts that satisfy the following rules:

(a) $is\_a(X, Y) \wedge part\_of(Y, Z) \rightarrow part\_of(X, Z)$
(b) $part\_of(X, Y) \wedge is\_a(Y, Z) \rightarrow part\_of(X, Z)$

The relations is_a, $part\_of$ correspond to the subsumption and the mereology relation, respectively, which are two of the most common relations encountered in KBs [43]. Recent studies have noted that many real world KB relations have very few facts [61], raising the importance of generalising with limited number of facts. To test our model in the presence of sparse long-tail relations, we kept the created datasets sufficiently small. For each type of the aforementioned rules, we extract 200 facts that satisfy them from Wikidata. We construct two datasets that we dub WD and WD_{++}. The dataset WD contains only the facts that satisfy rule (**a**). WD_{++} extends WD by also including the facts satisfying rule (**b**). The evaluation protocol was the following: For every dataset, we split all the facts randomly in train (80%), validation (10%), and test (10%) set, such that the validation and test sets only contain a subset of the rules' consequents in the form of $part\_of(X, Z)$. Table 1 provides details regarding the respective size of each dataset.

Table 1: Statistics of the experimental datasets.

Dataset	$\lvert$ **E** $\rvert$	$\lvert$ **R** $\rvert$	**#Train**	**#Valid**	**#Test**
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
WD	418	2	550	25	25
WD_{++}	763	2	1,120	40	40

4.2 Evaluation Protocol & Implementation Details

In the KBC task the models are evaluated based on their capability to answer queries such as $R(subject, \mathbf{?})$ and $R(\mathbf{?}, object)$ [10]; predicting, thus, the missing entity. Specifically, all the possible corruptions are obtained by replacing either the *subject* or the *object* and the entities are ranked based on the values of the implausibility score. The models should assign lower implausibility scores to valid facts and higher scores to implausible ones. We use the "**Filtered**" setting protocol [10], i.e., not taking any corrupted facts that exist in KB into account. We employ three common evaluation metrics: mean rank (MR), mean reciprocal rank (MRR), and Hits@10 (i.e., the proportion of the valid/test triples ranking in top 10 predictions). Higher MRR or higher Hits@10 indicate better performance. On the contrary, lower MR indicates better performance.

The reported results are given for the best set of hyperparameters evaluated on the validation set using grid search. Varying the batch size had no effect on the performance. Therefore, we divided every epoch into 10 mini-batches. The hyperparameter search space was the following: $\#_{negs_E} \in \{1, 2, 3, 4, 5, 8, 10, 12, 15\}$, $\#_{negs_R} \in \{0, 1, 2\}$, $\eta \in \{0.8, 0.5, 0.2, 0.1, 0.05, 0.01, 0.005\}$, $\beta \in \{\lfloor \frac{3n}{4} \rfloor, \lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{4} \rfloor, 0\}$, $\gamma \in \{7.0, 5.0, 2.0, 1.5, 1.0, 0.8, 0.5, 0.2, 0.1\}$, the embeddings' dimension $n \in \{40, 100, 200\}$, and $\lambda \in \{2.0, 1.5, 1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.0\}$. We used early stopping based on the validation's set filtered MRR performance, computed every 50 epochs with a maximum number of 2000 epochs. Due to space limitation, we report the best hyper-parameters in the Supplemental Material provided in [24].

4.3 Results & Analysis

Table 2 compares the experimental results of our HyperKG model with previous published results on WN18RR and FB15k-237 datasets. We have experimentally validated that both datasets present power-law degree distributions. Additionally, WN18RR contains more hierarchical-like relations compared to FB15k-237 [5]. We compare against the shallow KB embedding models DISTMULT [62], ComplEx

Table 2: Experimental results on WN18RR and FB15k-237 test sets. [⋆]: Results are taken from Nguyen et al. [33].

Method	**Type**	**WN18RR**			**FB15k-237**		
		MR	MRR	Hits@10	MR	MRR	Hits@10
DISTMULT [62] [⋆]	Bilinear	5110	0.43	0.49	254	0.24	0.41
ComplEx [51] [⋆]	Bilinear	5261	0.44	0.51	339	0.24	0.42
TransE [10] [⋆]	Translational	3384	0.22	0.50	347	0.29	0.46
HyperKG (Möbius addition)	Translational	4668	0.30	0.44	822	0.19	0.32
HyperKG (no regularisation)	Translational	5569	0.30	0.46	318	0.25	0.41
HyperKG	Translational	4165	0.41	0.50	272	0.28	0.45

[51] and TransE [10], which constitute important representatives of bilinear and translational models. We exclude from our comparison recent work that explores different types of training regimes such as adversarial training, the inclusion of reciprocal facts and/or multiple geometrical spaces [5, 11, 23, 26, 48] to make the analysis less biased to factors that could overshadow the importance of the embedding space. We give the results of our algorithm under the HyperKG listing. When we compare the performance of HyperKG and TransE on WN18RR, we see that HyperKG achieves almost the double MRR score. This shows that the lower MRR performance on certain datasets is not an intrinsic characteristic of the translational models, but a restriction that can be lifted by the right choice of geometrical space. On the WN18RR dataset, HyperKG exhibits slightly lower Hits@10 performance compared to ComplEx. Moreover, HyperKG achieves a better MR score compared to the bilinear models on WN18RR, but worse compared to TransE. On the FB15k-237 dataset, HyperKG and TransE demonstrate almost the same behaviour outperforming DISTMULT and ComplEx in terms of MRR and Hits@10. Since this performance gap is small, we hypothesise that this is due to a less fine-grained hyperparameter tuning. Interestingly, HyperKG achieves a better MR score compared to TransE on FB15k-237, but, still, worse compared to DISTMULT.

We also report in Table 2 two additional experiments where we explore the performance boost that our regularisation scheme brings as well as the behaviour of HyperKG when the Möbius addition is used instead of the Euclidean one. In the experiment where the Möbius addition was used, we removed the constraint for the entity vectors to have a norm less than 0.5. Although the Möbius addition is non-commutative, we found beneficial to keep the permutation matrix. Nonetheless, we do not use our regularisation scheme. Therefore, the implausibility score is $d_p(\boldsymbol{s} \boxplus \Pi_\beta \boldsymbol{o}, \boldsymbol{r})$. To investigate the effect of our proposed regularisation scheme, we show results where our regularisation scheme, defined in Equation (8), is not used, keeping, however, the rest of the architecture the same. Comparing the performance of the HyperKG variation using the Möbius addition against the performance of the HyperKG without regularisation, we can observe that we can achieve better results in terms of MRR and Hits@10 by using the Euclidean addition. This can be explained as follows. Generally, there is no unique and universal geometrical space adequate for every dataset [20]. To recover Euclidean Space from the Poincaré-ball model equipped with the Möbius addition, the ball's radius should grow to infinity [52]. Instead, by using the Euclidean addition and since the hyperbolic metric is locally Euclidean, HyperKG can model facts for which the Euclidean Space is more appropriate by learning to retain small distances. Last but not least, we can observe that our proposed regularisation scheme is beneficial in terms of MR, MRR and Hits@10 on both datasets. Overall, the hyperbolic space appears more beneficial for datasets that contain many hierarchical-like relations such as WN18RR, without a significant performance degradation in the other case.

Table 3 reports the results on the WD and WD_{++} datasets. We compare HyperKG performance against that of TransE and ComplEx. It can be observed that none of the models manages to totally capture the statistical regularities of these datasets. All the models undergo similar Hits@10 performance on both datasets. HyperKG and TransE, that both have convex relation spaces, outperform ComplEx on both datasets in terms of MRR and Hits@10. Furthermore, the translational models show a relatively steady performance compared to ComplEx, whose performance deteriorates in the presence of the two rules appearing in WD_{++}. With regard to MR, HyperKG closes the gap between translational and bilinear models on WD and shows the best performance on WD_{++}. Our results point to a promising direction for developing less expressive KB embedding models which can, however, better represent certain types of rules.

Table 3: Experimental results on WD and WD_{++} test sets.

Method	**WD**			$\textbf{WD}_{++}$		
	MR	MRR	Hits@10	MR	MRR	Hits@10
ComplEx	1.22	0.92	0.98	2.42	0.81	0.92
TransE	2.52	0.88	0.96	2.01	0.89	0.98
HyperKG	1.32	0.98	0.98	1.36	0.93	0.98

5 Conclusion and Outlook

In this paper, we examined the importance of the geometrical space for the task of KBC. We showed that the lagging performance of translational models compared to the bilinear ones is not an intrinsic characteristic of them but a restriction that can be lifted in the hyperbolic space. Our results validated that the right choice of geometrical space is a critical decision that impacts the performance of KB embedding models. Our findings also shed light on understanding which KBs mostly benefit from the use of hyperbolic embeddings. Moreover, we demonstrated a new promising direction for developing models that, although not fully expressive, allow to better represent certain families of rules; opening up for more fine-grained reasoning tasks. In the future, we plan to extend our approach to the bilinear family of models.

Acknowledgments We would like to thank the anonymous reviewers for their insightful comments on the paper.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of databases: the logical level. Addison-Wesley Longman Publishing Co., Inc. (1995)
2. Ahlfors, L.V.: Invariant operators and integral representations in hyperbolic space. Mathematica Scandinavica 36(1), 27–43 (1975)
3. Alstott, J., Bullmore, E., Plenz, D.: powerlaw: a python package for analysis of heavy-tailed distributions. PloS one 9(1), e85777 (2014)
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics 25(1), 25 (2000)
5. Balažević, I., Allen, C., Hospedales, T.: Multi-relational poincaré graph embeddings. In: NeurIPS (2019)
6. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
7. Boguná, M., Papadopoulos, F., Krioukov, D.: Sustaining the internet with hyperbolic mapping. Nature communications 1, 62 (2010)
8. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD (2008)
9. Bonnabel, S.: Stochastic gradient descent on riemannian manifolds. IEEE Trans. Automat. Contr. 58(9), 2217–2229 (2013)
10. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NeurIPS (2013)
11. Cai, L., Wang, W.Y.: KBGAN: Adversarial learning for knowledge graph embeddings. In: NAACL (Jun 2018), `https://www.aclweb.org/anthology/N18-1133`
12. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: AAAI (2018)
13. Ebisu, T., Ichise, R.: Toruse: Knowledge graph embedding on a lie group. In: AAAI (2018)
14. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing wikidata to the linked data web. In: ISWC (2014)
15. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: ACM SIGCOMM computer communication review. vol. 29, pp. 251–262. ACM (1999)
16. Feng, J., Huang, M., Wang, M., Zhou, M., Hao, Y., Zhu, X.: Knowledge graph embedding by flexible translation. In: KR (2016)
17. Ganea, O., Becigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: ICML. pp. 1646–1655 (2018)
18. Getoor, L., Taskar, B.: Introduction to statistical relational learning, vol. 1. MIT press Cambridge (2007)
19. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. pp. 249–256 (2010)

20. Gu, A., Sala, F., Gunel, B., Ré, C.: Learning mixed-curvature representations in product spaces. In: ICLR (2018)
21. Gutiérrez-Basulto, V., Schockaert, S.: From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. In: KR (2018)
22. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: ACL-IJCNLP. pp. 687–696 (2015)
23. Kazemi, S.M., Poole, D.: Simple embedding for link prediction in knowledge graphs. In: NeurIPS. pp. 4284–4295 (2018)
24. Kolyvakis, P., Kalousis, A., Kiritsis, D.: HyperKG: Hyperbolic knowledge graph embeddings for knowledge base completion. arXiv preprint arXiv:1908.04895 (2019)
25. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguñá, M.: Hyperbolic geometry of complex networks. Phys. Rev. E 82, 036106 (Sep 2010)
26. Lacroix, T., Usunier, N., Obozinski, G.: Canonical tensor decomposition for knowledge base completion. In: ICML (2018)
27. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web 6(2), 167–195 (2015)
28. Li, M., Jia, Y., Wang, Y., Li, J., Cheng, X.: Hierarchy-based link prediction in knowledge graphs. In: WWW (2016), https://doi.org/10.1145/2872518.2889387
29. Miller, G.: WordNet: An electronic lexical database. MIT press (1998)
30. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: Proceedings of ACL-08: HLT (2008), http://aclweb.org/anthology/P08-1028
31. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. The Journal of Logic Programming 19, 629–679 (1994)
32. Newman, M.E.: Power laws, pareto distributions and zipf's law. Contemporary physics 46(5), 323–351 (2005)
33. Nguyen, D.Q., Nguyen, T.D., Nguyen, D.Q., Phung, D.: A novel embedding model for knowledge base completion based on convolutional neural network. In: NAACL (2018), http://aclweb.org/anthology/N18-2053
34. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. Proceedings of the IEEE 104(1), 11–33 (2016)
35. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. In: AAAI (2016), http://dl.acm.org/citation.cfm?id=3016100.3016172
36. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: ICML (2011)
37. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. In: NeurIPS (2017)
38. Nickel, M., Kiela, D.: Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In: ICML (2018)
39. Papadopoulos, F., Aldecoa, R., Krioukov, D.: Network geometry inference using common neighbors. Physical Review E 92(2), 022807 (2015)
40. Rassias, T.M., Suksumran, T.: An inequality related to möbius transformations. arXiv preprint arXiv:1902.05003 (2019)
41. Richardson, M., Domingos, P.: Markov logic networks. Machine learning 62(1-2), 107–136 (2006)
42. Sala, F., De Sa, C., Gu, A., Re, C.: Representation tradeoffs for hyperbolic embeddings. In: ICML (2018), http://proceedings.mlr.press/v80/sala18a.html
43. Schwarz, U., Smith, B.: Ontological relations. Applied Ontology. An Introduction 219, 234 (2008)

44. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: NeurIPS. pp. 926–934 (2013)
45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15, 1929–1958 (2014)
46. Steyvers, M., Tenenbaum, J.B.: The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science 29(1), 41–78 (2005)
47. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW (2007)
48. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: ICLR (2019)
49. Suzuki, A., Enokida, Y., Yamanishi, K.: Riemannian transe: Multi-relational graph embedding in non-euclidean space (2019), https://openreview.net/forum?id=r1xRW3A9YX
50. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference. In: Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality (2015), http://aclweb.org/anthology/W15-4007
51. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: ICML (2016)
52. Ungar, A.A.: Beyond the Einstein addition law and its gyroscopic Thomas precession: The theory of gyrogroups and gyrovector spaces, vol. 117. Springer Science & Business Media (2012)
53. Van Der Hofstad, R.: Random graphs and complex networks. Available on http://www. win. tue. nl/rhofstad/NotesRGCN. pdf 11 (2009)
54. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Commun. ACM 57(10), 78–85 (Sep 2014), http://doi.acm.org/10.1145/2629489
55. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering 29(12), 2724–2743 (2017)
56. Wang, Y., Gemulla, R., Li, H.: On multi-relational link prediction with bilinear models. In: AAAI (2018)
57. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI (2014)
58. West, R., Gabrilovich, E., Murphy, K., Sun, S., Gupta, R., Lin, D.: Knowledge base completion via search-based question answering. In: WWW (2014)
59. Xiao, H., Huang, M., Zhu, X.: From one point to a manifold: Knowledge graph embedding for precise link prediction. In: IJCAI (2016)
60. Xie, Q., Ma, X., Dai, Z., Hovy, E.: An interpretable knowledge transfer model for knowledge base completion. In: ACL (2017)
61. Xiong, W., Yu, M., Chang, S., Guo, X., Wang, W.Y.: One-shot relational learning for knowledge graphs. In: EMNLP (Oct-Nov 2018)
62. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: ICLR (2015)
63. Zipf, G.K.: Human Behaviour and the Principle of Least Effort. Addison-Wesley (1949)