Holistic Representations for Memorization and Inference

Yunpu Ma[*]
LMU
Siemens AG
Otto-Hahn-Ring 6
81739 Munich

Marcel Hildebrandt
LMU
Siemens AG
Otto-Hahn-Ring 6
81739 Munich

Stephan Baier
LMU
Oettingenstr. 67
80538 Munich

Volker Tresp
LMU
Siemens AG
Otto-Hahn-Ring 6
81739 Munich

Abstract

In this paper we introduce a novel holographic memory model for the distributed storage of complex association patterns and apply it to knowledge graphs. In a knowledge graph, a labelled link connects a subject node with an object node, jointly forming a subject-predicate-objects triple. In the presented work, nodes and links have initial random representations, plus *holistic representations* derived from the initial representations of nodes and links in their local neighbourhoods. A memory trace is represented in the same vector space as the holistic representations themselves. To reduce the interference between stored information, it is required that the initial random vectors should be pairwise quasi-orthogonal. We show that pairwise quasi-orthogonality can be improved by drawing vectors from heavy-tailed distributions, e.g., a Cauchy distribution, and, thus, memory capacity of holistic representations can significantly be improved. Furthermore, we show that, in combination with a simple neural network, the presented holistic representation approach is superior to other methods for link predictions on knowledge graphs.

1 INTRODUCTION

An associative memory is a key concept in artificial intelligence and cognitive neuroscience for learning and memorizing relationships between entities and concepts. Various computational models of associative memory have been proposed, see, e.g., [Hopfield 1982; Gentner 1983]. One important family of associative memory models is the holographic associative memory (HAM), which was first proposed in [Gabor 1969]. HAMs can store a large number of stimulus-response pairs as additive superpositions of memory traces. It has been suggested that this holographic storage is related to the working principle of the human brain [Westlake 1970].

An important extension to the HAM is based on holographic reduced representations (HRR) [Plate 1995]. In HRR, each entity or symbol is represented as a vector defined in a continuous space. Associations between two entities are compressed in the same vector space via a vector binding operation; the resulting vector is a memory trace. Two associated entities are referred to as a *cue-filler* pair, since a noisy version of the *filler* can be recovered from the memory trace and the *cue* vector via a decoding operation. Multiple *cue-filler* pairs can be compressed in a single memory trace through superposition. Associations can be read out from this single trace, however with large distortions. Thus, a clean-up mechanism was introduced into HRR, such that associations can be *retrieved* with high probability.

The number of associations which can be compressed in a single trace is referred to as *memory capacity*. It has been shown in [Plate 1995] that the memory capacity of the HRR depends on the degree of the pairwise orthogonality of initial random vectors associated with the entities.

Quasi-orthogonality was put forward in [Diaconis et al. 1984; Hall et al. 2005]. They informally stated that “most independent high-dimensional random vectors are nearly orthogonal to each other”. A rigorous mathematical justification to this statement has only recently been given in [Cai et al. 2012; Cai et al. 2013], where the density function of pairwise angles among a large number of Gaussian random vectors was derived. To the best of our knowledge, density functions for other distributions have not been derived, so far. As a first contribution, we will derive a significantly improved quasi-orthogonality, and

[*] yunpu.ma@siemens.com

we show that memory capacity of holographic representations can significantly be improved. Our result could potentially have numerous applications, e.g., in sparse random projections or random geometric graphs [Penrose 2003].

After the HRR had been proposed, it had mainly been tested on small toy datasets. Quasi-orthogonality becomes exceedingly important when a large amount of entities needs to be initialized with random vectors, as in applications involving large-scale knowledge graphs.

Modern knowledge graphs (KGs), such as FREEBASE [Bollacker et al. 2008], YAGO [Suchanek et al. 2007], and GDELT [Leetaru et al. 2013], are relational knowledge bases, where nodes represent entities and directed labelled links represent predicates. An existing labelled link between a head node (or subject) and a tail node (or object) is a triple and represents a fact, e.g. (*California, locatedIn, USA*).

As a second contribution, we demonstrate how the holographic representations can be applied to KGs. First, one needs to define association pairs (or *cue-filler* pairs). We propose that the representation of a *subject* should encode all *predicate-object* pairs, such that given the *predicate* representation as a *cue*, the *object* should be recovered or at least recognized. Similarly, the representation of an *object* should encode all *predicate-subject* pairs, such that the *subject* can be retrieved after decoding with the *predicate* representation. We call those representations *holistic*, since they are inspired by the semantic holism in the philosophy of language, in the sense that an abstract entity can only be comprehended through its relationships to other abstract entities.

So far we have discussed memory formation and memory retrieval. Another important function is the generalization of stored memory to novel facts. This has technical applications and there are interesting links to human memory. From a cognitive neuroscientist point of view, the brain requires a dual learning system: one is the hippocampus for rapid memorization, and the other is the neocortex for gradual consolidation and comprehension. This hypothesis is the basis for the *Complementary Learning System* (CLS) which was first proposed in [McClelland et al. 1995]. Connections between KGs and long-term declarative memories has recently been stated in [Tresp et al. 2017a; Ma et al. 2018; Tresp et al. 2017b].

As a third contribution of this paper, we propose a model which not only memorizes patterns in the training datasets through holistic representations, but also is able to infer missing links in the KG, by a simple neural network that uses the holistic representations as input representations. Thus, our model realizes a form of a *complementary learning system*. We compare our results on multiple datasets with other state-of-the-art link prediction models, such as RESCAL [Nickel et al. 2011; Nickel et al. 2012], DISTMULT [Yang et al. 2014], COMPLEX [Trouillon et al. 2016], and R-GCN [Schlichtkrull et al. 2018].

The above mentioned learning-based methods model the KGs by optimizing the latent representaions of entities and predicates through minimizing the loss function. It had been observed that latent embeddings are suitable for capturing global connectivity patterns and generalization [Nickel et al. 2016a; Toutanova et al. 2015], but are not as good in memorizing unusual patterns, such as patterns associated with locally and sparsely connected entities. This motivates us to *separate* the memorization and inference tasks. As we will show in our experiments, our approach can, on the one hand, memorize local graph structures, but, on the other hand, also generalizes well to global connectivity patterns, as required by complementary learning systems.

Note, that in our approach holistic representations are derived from random vectors and are **not** learned from data via backpropagation, as in most learning-based approaches to representation learning on knowledge graphs. One might consider representations derived from random vectors to be biologically more plausible, if compared to representations which are learned via complex gradient based update rules [Nickel et al. 2016a]. Thus, in addition to its very competitive technical performance, one of the interesting aspects of our approach is its biological plausibility.

In Section 2 we introduce notations for KGs and embedding learning. In Section 3 we discuss improved quasi-orthogonality by using heavy-tailed distributions. In Section 4 we propose our own algorithm for holistic representations, and test it on various datasets. We also discuss how the memory capacity can be improved. In Section 5 we propose a model which can infer implicit links on KGs through holistic representations. Section 6 contains our conclusions.

2 REPRESENTATION LEARNING

In this section we provide a brief introduction to representation learning in KGs, where we adapt the notation of [Nickel et al. 2016b]. Let $\mathcal{E}$ denotes the set of entities, and $\mathcal{P}$ the set of predicates. Let N_e be the number of entities in $\mathcal{E}$, and N_p the number of predicates in $\mathcal{P}$.

Given a predicate $p \in \mathcal{P}$, the characteristic function $\phi_p : \mathcal{E} \times \mathcal{E} \to \{1, 0\}$ indicates whether a triple $(\cdot, p, \cdot)$ is true or false. Moreover, $\mathcal{R}_p$ denotes the set of all subject-object pairs, such that $\phi_p = 1$. The entire KG can be written as $\chi = \{(i,j,k)\}$, with $i = 1, \cdots, N_e$, $j = 1, \cdots, N_p$, and $k = 1, \cdots, N_e$.

We assume that each entity and predicate has a unique latent representation. Let $\mathbf{a}_{e_i}$, $i = 1, \cdots, N_e$, be the representations of entities, and $\mathbf{a}_{p_i}$, $i = 1, \cdots, N_p$, be the representations of predicates. Note that $\mathbf{a}_{e_i}$ and $\mathbf{a}_{p_i}$ could be real- or complex-valued vectors/matrices.

A probabilistic model for the KG χ is defined as $\Pr(\phi_p(s,o) = 1|\mathcal{A}) = \sigma(\eta_{spo})$ for all (s,p,o)-triples in χ, where $\mathcal{A} = \{\mathbf{a}_{e_i}\}_i^{N_e} \cup \{\mathbf{a}_{p_i}\}_i^{N_p}$ denotes the collection of all embeddings; $\sigma(\cdot)$ denotes the sigmoid function; and η_{spo} is the a function of latent representations, $\mathbf{a}_s$, $\mathbf{a}_p$ and $\mathbf{a}_o$. Given a labeled dataset containing both true and false triples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, with $x_i \in \chi$, and $y_i \in \{1, 0\}$, latent representations can be learned. Commonly, one minimizes a binary cross-entropy loss

$$-\frac{1}{m}\sum_{i=1}^{m}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right) + \lambda||\mathcal{A}||_2^2, \tag{1}$$

where m is the number of training samples, and λ is the regularization parameter; $p_i := \sigma(\eta_{x_i})$ with $\sigma(\cdot)$ being the sigmoid function. η_{spo} is defined differently in various models.

For instance, for RESCAL entities are represented as r-dimensional vectors, $\mathbf{a}_{e_i} \in \mathbb{R}^r$, $i = 1, \cdots, N_e$, and predicates are represented as matrices, $\mathbf{a}_{p_i} \in \mathbb{R}^{r \times r}$, $i = 1, \cdots, N_p$. Moreover, one uses $\eta_{spo} = \mathbf{a}_s^\intercal \mathbf{a}_p \mathbf{a}_o$.

For DISTMULT, $\mathbf{a}_{e_i}, \mathbf{a}_{p_j} \in \mathbb{R}^r$, with $i = 1, \cdots, N_e$, $j = 1, \cdots, N_p$; $\eta_{spo} = \langle \mathbf{a}_s, \mathbf{a}_p, \mathbf{a}_o \rangle$, where $\langle \cdot, \cdot, \cdot \rangle$ denotes the tri-linear dot product.

For COMPLEX, $\mathbf{a}_{e_i}, \mathbf{a}_{p_j} \in \mathbb{C}^r$, with $i = 1, \cdots, N_e$, $j = 1, \cdots, N_p$; $\eta_{spo} = \Re(\langle \mathbf{a}_s, \mathbf{a}_p, \bar{\mathbf{a}}_o \rangle)$, where the bar denotes complex conjugate, and $\Re$ denotes the real part.

3 DERIVATION OF ϵ-ORTHOGONALITY

As we have discussed in the introduction, quasi-orthogonality of the random vectors representing the entities and the predicates is required for low interference memory retrieval. In this section we investigates the asymptotic distribution of pairwise angles in a set of independently and identically drawn random vectors. In particular, we study random vectors drawn from either a Gaussian or a heavy-tailed Cauchy distribution distribution. A brief summary of notations is referred to the A.7. First we define the term "ϵ-orthogonality".

Definition 1. *A set of n vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is said to be pairwise ϵ-orthogonal, if $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| < \epsilon$ for $i, j = 1, \cdots, n$, $i \neq j$.*

Here, $\epsilon > 0$ is a small positive number, and $\langle \cdot, \cdot \rangle$ denotes the inner product in the vector space.

3.1 ϵ-ORTHOGONALITY FOR A GAUSSIAN DISTRIBUTION

In this section we revisit the empirical distribution of pairwise angles among a set of random vectors. More specifically, let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent q-dimensional Gaussian variables with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$. Denote with Θ_{ij} the angle between $\mathbf{X}_i$ and $\mathbf{X}_j$, and $\rho_{ij} := \cos\Theta_{ij} \in [-1, 1]$. [Cai et al. 2012; Muirhead 2009] derived the density function of ρ_{ij} in the following Lemma.

Lemma 1. *Consider ρ_{ij} as defined above. Then $\{\rho_{ij} | 1 < i < j \leq n\}$ are pairwise i.i.d. random variables with the following asymptotic probability density function*

$$g(\rho_{\mathrm{G}}) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} (1 - \rho_{\mathrm{G}}^2)^{\frac{q-3}{2}}, \quad |\rho_{\mathrm{G}}| < 1, \tag{2}$$

with fixed dimensionality q.

[Cai et al. 2013] also derived the following Theorem 1.

Theorem 1. *Let the empirical distribution μ_n of pairwise angles $\Theta_{ij}, 1 \leq i < j \leq n$ be defined as $\mu_n := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \delta_{\Theta_{ij}}$. With fixed dimension q, as $n \to \infty$, μ_n converges weakly to the distribution with density*

$$h(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} (\sin\theta)^{q-2}, \quad \theta \in [0, \pi]. \tag{3}$$

From the above distribution function we can derive the upper bound of quasi-orthogonal random vectors with pairwise ϵ-orthogonality in the Euclidean space $\mathbb{R}^q$.

Corollary 1. *Consider a set of independent q-dimensional Gaussian random vectors which are pairwise ϵ-orthogonal with probability $1-\nu$, then the number of such Gaussian random vectors is bounded by*

$$N \leq \sqrt[4]{\frac{\pi}{2q}}\, \mathrm{e}^{\frac{\epsilon^2 q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \tag{4}$$

The derivation is given in A.1. Due to the symmetry of density function $g(\rho_{\mathrm{G}})$, we immediately have $\mathbb{E}[\rho_{\mathrm{G}}] = 0$, moreover, $\mathbb{E}[\theta] = \frac{\pi}{2}$. However, for the later use, it is important to consider the expected absolute value of ρ_{G}:

Corollary 2. *Consider a set of n q-dimensional random Gaussian vectors, we have*

$$\lambda_G := \mathbb{E}[|\rho_{\mathrm{G}}|] = \sqrt{\frac{2}{\pi q}}. \tag{5}$$

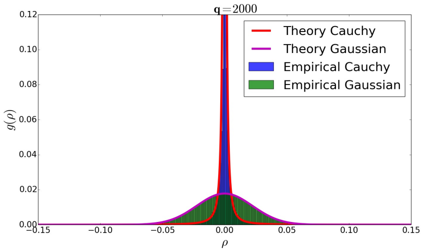


Figure 1: Empirical pairwise angle distribution in a set of Gaussian random vectors (green) is compared with theoretical prediction Eq. 2 (magenta); Empirical pairwise angle distribution in a set of Cauchy random vectors (blue) is compared with prediction Eq. 6 (red)

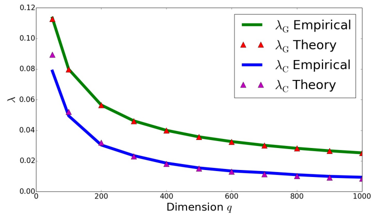


Figure 2: Compare λ_{G} and λ_{C} from simulation and theory, see Eq. 5 and Eq. 9.

Note, that the quantity $\frac{\pi}{2} - \arccos \mathbb{E}[|\rho_{\mathrm{G}}|]$ has a clear geometrical meaning: It indicates the expected deviation from $\frac{\pi}{2}$ of pairwise angles. In fact, in the extreme case when $q \rightarrow \infty$, the deviation converges to 0 with the rate $\sqrt{q}$.

3.2 ϵ-ORTHOGONALITY FOR A CAUCHY DISTRIBUTION

In this subsection, we show that the set of random vectors whose elements are initialized with a heavy-tailed distribution, e.g., a Cauchy distribution $\mathcal{C}\,(0,1)$, has improved ϵ-orthogonality. The intuition is as follows: Consider a set of q-dimensional random vectors initialized with a heavy-tailed distribution. After normalization, each random vector can be approximated by only the elements which significantly deviate from zero and were drawn from the heavy tails. If the number of those elements is k with $k \ll q$, then there are at most $\binom{q}{k}$ orthogonal random vectors.

Moreover, $\binom{q}{k} \approx \frac{q^k}{k\Gamma(k)}$ could be much larger than $\sqrt[4]{\frac{\pi}{2q}} \mathrm{e}^{\frac{\epsilon^2 q}{4}}$ from Eq. 4, when q is sufficiently large, $k \ll q$, and $\epsilon \rightarrow 0$. In other words, under stricter quasi-orthogonality condition with smaller ϵ, random vectors drawn from a heavy-tailed distribution could have more pairs satisfying the quasi-orthogonality condition.

Consider a set of q-dimensional Cauchy random vectors. As $q \rightarrow \infty$ the approximate density function of ρ_{ij}, with $1 \leq i < j \leq n$ is described in the following conjecture.

Conjecture 1. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent q-dimensional random vectors whose elements are independently and identically drawn from Cauchy a distribution $\mathcal{C}(0,1)$. Moreover, consider the angle Θ_{ij} between $\mathbf{X}_i$, and $\mathbf{X}_j$. Then, as $q \rightarrow \infty$, $\rho_{ij} := \cos \Theta_{ij} \in [-1,1]$, $1 \leq i < j \leq n$ are pairwise i.i.d. with a density function approximated by*

$$g(\rho_{\mathrm{C}}) = -\frac{2}{\pi^2 q^2 \rho_{\mathrm{C}}^3} \cdot \frac{1}{z^{\frac{3}{2}}} \left[\mathrm{e}^{\frac{1}{\pi z}} \operatorname{Ei}\left(-\frac{1}{\pi z}\right)\right], \tag{6}$$

where $z := \frac{1}{q^2}\left(\frac{1}{\rho_{\mathrm{C}}^2} - 1\right)$, and the exponential integral $\operatorname{Ei}(x)$ is defined as $\operatorname{Ei}(x) = -\int\limits_{-x}^{\infty} \frac{\mathrm{e}^{-t}}{t} dt$.

The intuition behind the conjecture is as follows. Suppose $\mathbf{X} = (X_1, \cdots, X_q)$ and $\mathbf{Y} = (Y_1, \cdots, Y_q)$ are random vector variables, and assume that elements of $\mathbf{X}$ and $\mathbf{Y}$ are independently Gaussian distributed. In order to derive $g(\rho_{\mathbf{X},\mathbf{Y}})$ in Lemma 1, [Cai et al. 2012; Muirhead 2009] compute the distribution function for $\frac{\boldsymbol{\alpha}^\intercal \cdot \mathbf{X}}{||\mathbf{X}||}$ instead, where $\boldsymbol{\alpha}^\intercal \boldsymbol{\alpha} = 1$. In particular, they assume that $\boldsymbol{\alpha} = (1, 0, \cdots, 0)$. The underlying reason for this assumption is that the random vector $\frac{\mathbf{X}}{||\mathbf{X}||}$ is uniformly distributed on the $(q-1)$-dimensional sphere.

Here, elements of $\mathbf{X}$ and $\mathbf{Y}$ are independently Cauchy distributed. We derive the approximation in Eq. 6 under the same assumption by taking $g(\rho_{\mathbf{X},\mathbf{Y}}) \approx \frac{X_1}{\sqrt{X_1^2 + \cdots + X_q^2}}$. Furthermore, we introduce a new variable $z_{\mathbf{X},\mathbf{Y}} := \frac{1}{q^2}\left(\frac{1}{\rho_{\mathbf{X},\mathbf{Y}}^2} - 1\right) = \frac{1}{q^2} \frac{X_2^2 + \cdots + X_q^2}{X_1^2}$, and derive the density function $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$ by using the generalized central limit theorem [Gnedenko et al. 1954] and properties of quotient distributions of two independent random variables. $g(\rho_{\mathbf{X},\mathbf{Y}})$ can be directly obtained from $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$ by a variable transform. More details and derivation are referred to the A.2.

We turn to study the limiting behaviour of the density function when ρ approaches zero. In this case, the variable z defined in in Conjecture 1 can be approximated by $z \approx \frac{1}{q^2 \rho_{\mathrm{C}}^2}$. Using properties of the exponential integral, as $q \rightarrow \infty$, the density function in Eq. 6 can be approximated by its Laurent series,

$$g(\rho_{\mathrm{C}}) \approx \frac{2}{\pi q \rho_{\mathrm{C}}^2} - \frac{2}{q^3 \rho_{\mathrm{C}}^4} + \frac{4\pi}{q^5 \rho_{\mathrm{C}}^6} + \mathcal{O}\left(\frac{1}{q^7 \rho_{\mathrm{C}}^8}\right) \tag{7}$$

In the following corollary we give the upper bound of the number of pairwise ϵ-orthogonal Cauchy random vectors using Eq. 6.

Corollary 3. *Consider a set of independent q-dimensional Cauchy random vectors which are pairwise ϵ-orthogonal with probability $1-\nu$, then the number of such Cauchy random vectors is bounded by*

$$N \leq \sqrt{\frac{\pi\epsilon q}{4}}\left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \qquad (8)$$

Let us compare the prefactor of this upper bound for two distributions: That is $\sqrt[4]{\frac{\pi}{2q}}\,\mathrm{e}^{\frac{\epsilon^2 q}{4}}$ for the Gaussian distribution, and $\sqrt{\frac{\pi\epsilon q}{4}}$ for the Cauchy distribution. Under strict quasi-orthogonal conditions with arbitrarily small but fixed $\epsilon > 0$, for the dimension $q \gg 2\sqrt[3]{\frac{1}{\pi\epsilon^2}}$ we have that $\sqrt{\frac{\pi\epsilon q}{4}} \gg \sqrt[4]{\frac{\pi}{2q}}\,\mathrm{e}^{\frac{\epsilon^2 q}{4}} \approx \sqrt[4]{\frac{\pi}{2q}}$. It implies that in sufficiently high-dimensional spaces, random vectors which are independently drawn from a Cauchy distribution are more likely to satisfy the pairwise ϵ-orthogonality condition - particularly when $\epsilon \ll 1$.

Remark 1. *For the later use, we define λ_{C} as $\lambda_{\mathrm{C}} := \mathbb{E}[|\rho_{\mathrm{C}}|]$ for the case of Cauchy distribution. However, no simple analytic form is known for this integral. Thus we use the following numerically stable and non-divergent equation to approximate λ_{C},*

$$\lambda_{\mathrm{C}} \approx -\frac{4q}{\pi^2}\int_0^1 \rho\left[\mathrm{e}^{\frac{q^2\rho^2}{\pi}}\,\mathrm{Ei}\left(-\frac{q^2\rho^2}{\pi}\right)\right]\mathrm{d}\rho. \qquad (9)$$

This simpler form is derived from Eq. 6 using the approximation $z \approx \frac{1}{q^2\rho^2}$.

Fig. 1 shows the empirical distribution of ρ_{G} in a set of Gaussian random vectors (green) compared with theoretical prediction in Eq.2 (magenta); and the empirical distribution of ρ_{C} in a set of Cauchy random vectors (blue) compared with theoretical prediction (red). In the case of Cauchy random vectors, the leading orders of the Laurent expansion of Eq. 6 are used, see Eq. 7. For the empirical simulation, 10000 random vectors with dimensionality $q = 2000$ were drawn independently from either a Gaussian or a Cauchy distribution.

In addition, in Fig. 2 we plot λ_{G} and λ_{C} as a function of q in comparison with the theoretical predictions from Eq. 5 and Eq. 9, respectively, under the same simulation condition. It is necessary to emphasize that $\lambda_{\mathrm{C}}(q) < \lambda_{\mathrm{G}}(q)$ for all the dimensions q; this fact will be used to explain the relatively high memory capacity encoded from Cauchy random vectors.

In the Appendix, see Remark A 2, the distribution of elements from the normalized random variable $\frac{\mathbf{X}}{||\mathbf{X}||}$ is also considered. In particular, for normalized Cauchy random vector most of its elements are nearly zero, and it realizes a **sparse** representation.

4 HOLISTIC REPRESENTATIONS FOR KGS

4.1 HRR MODEL

First, we briefly review HRR. Three operations are defined in HRR to model associative memories: *encoding*, *decoding*, and *composition*.

Let $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$ be random vectors representing different entities. The encoding phase stores the association between $\mathbf{a}$ and $\mathbf{b}$ in a memory trace $\mathbf{a} * \mathbf{b}$, where $* : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^q$ denotes circular convolution, which is defined as $[\mathbf{a} * \mathbf{b}]_k = \sum_{i=0}^{q-1} a_i b_{(k-i) \bmod q}$.

A noisy version of $\mathbf{b}$ can be retrieved from the memory trace, using the item $\mathbf{a}$ as a cue, with: $\mathbf{b} \approx \mathbf{a} \star (\mathbf{a} * \mathbf{b})$, where $\star : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^q$ denotes the circular correlation [1]. It is defined as $[\mathbf{a} \star \mathbf{b}]_k = \sum_{i=0}^{q-1} a_i b_{(k+i) \bmod q}$. In addition, several associations can be superimposed in a single trace via the addition operation: $(\mathbf{a} * \mathbf{b}) + (\mathbf{c} * \mathbf{d}) + \cdots$.

4.2 HOLISTIC MODEL

Initially, each entity and predicate in a KG is associated with a q-dimensional normalized random vector, which is then normalized. We denote them as $\mathbf{r}_{e_i}^{\mathrm{G/C}}$, $i = 1, \cdots, N_e$, and $\mathbf{r}_{p_i}^{\mathrm{G/C}}$, $i = 1, \cdots, N_p$, respectively. The superscript indicates from which distribution vector elements are independently drawn, either the Gaussian or Cauchy distribution. If there is no confusion, we may omit the superscript.

Consider an entity e_i. Let $\mathcal{S}^s(e_i) = \{(p, o)|\phi_p(e_i, o) = 1\}$ be the set of all predicate-object pairs for which triples (e_i, p, o) is true and where e_i is the subject. We store these multiple associations in a single memory trace via circular correlation and superposition:

$$\mathbf{h}_{e_i}^s = \sum_{(p,o)\in\mathcal{S}^s(e_i)} \left[\mathrm{Norm}(\mathbf{r}_p \star \mathbf{r}_o) + \xi\mathbf{r}_{e_i}\right], \qquad (10)$$

where $\mathrm{Norm} : \mathbb{R}^q \to \mathbb{R}^q$ represents the normalization operation [2], which is defined as $\mathrm{Norm}(\mathbf{r}) := \frac{\mathbf{r}}{||\mathbf{r}||}$. Moreover, the hyper-parameter $\xi > 0$ determines the contribution of the individual initial representation $\mathbf{r}$.

[1] It uses the fact that $\mathbf{a} \star \mathbf{a} \approx \delta$, where δ is the identity operation of convolution.

[2] In other sections, we may obviate Norm operator in the equation for the sake of simplicity, since it can be shown that the circular correlation of two normalized high-dimensional random vectors are almost normalized.

Note, that the same entity e_i could also play the role of an object. For instance, the entity *California* could be the subject in the triple (*California, locatedIn, USA*), or the object in another triple (*Paul, livesIn, California*). Thus, it is necessary to have another representation to specify its role in the triples. Consider the set of subject-predicate pairs $\mathcal{S}^o(e_i) = \{(s,p)|\phi_p(s,e_i) = 1\}$ for which triples (s,p,e_i) are true. These pairs are stored in a single trace via $\mathbf{h}^o_{e_i} = \sum_{(s,p)\in\mathcal{S}^o(e_i)} [\mathrm{Norm}(\mathbf{r}_p \star \mathbf{r}_s) + \xi \mathbf{r}_{e_i}]$, where $\mathbf{h}^o_{e_i}$ is the representation of the entity e_i when it acts as an object.

For the later generalization task, the overall holistic representation for the entity e_i is defined as the summation of both representations, namely

$$\mathbf{h}_{e_i} = \mathbf{h}^s_{e_i} + \mathbf{h}^o_{e_i}. \tag{11}$$

In this way, the complete neighbourhood information of an entity can be used for generalization.

Furthermore, given a predicate p_i, the holistic representation $\mathbf{h}_{p_i}$ encodes all the subject-object pairs in the set $\mathcal{S}(p_i) = \{(s,o)|\phi_{p_i}(s,o) = 1\}$ via

$$\mathbf{h}_{p_i} = \sum_{(s,o)\in\mathcal{S}(p_i)} \left[\mathrm{Norm}(\mathbf{r}_s \star \mathbf{r}_o) + \xi \mathbf{r}_{p_i}\right]. \tag{12}$$

After storing all the association pairs into holistic features of entities and predicates, the initial randomly assigned representations are not required anymore and can be deleted. These representations are then fixed and not trainable unlike other embedding models.

After encoding, entity retrieval is performed via a circular convolution. Consider a concrete triple (e_1, p_1, e_2) with unknown e_2. The identity of e_2 could be revealed with the holistic representation of p_1 and the holistic representation of e_1 as a subject, namely $\mathbf{h}_{p_1}$ and $\mathbf{h}^s_{e_1}$. Then retrieval is performed as $\mathbf{h}_{p_1} * \mathbf{h}^s_{e_1}$. The associations can be retrieved from the holography memory with low fidelity due to interference. Therefore, after decoding, a clean-up operation is employed, as in the HRR model. Specifically, a nearest neighbour is determined using cosine similarity. The pseudo-code for encoding holistic representations is provided in A.6.

4.3 EXPERIMENTS ON MEMORIZATION

We test the memorization of complex structure on different datasets and compare the performance of different models. Recall that $\mathcal{R}_\mathrm{p}$ is the set of all true triples with respect to a given predicate p. Consider a possible triple $(\mathrm{s}, \mathrm{p}, \mathrm{o}) \in \mathcal{R}_\mathrm{p}$. The task is now to retrieve the object entity from holistic vectors $\mathbf{h}_\mathrm{s}$ and $\mathbf{h}_\mathrm{p}$, and to retrieve the subject entity from holistic vectors $\mathbf{h}_\mathrm{p}$ and $\mathbf{h}_\mathrm{o}$. As discussed, in retrieval, the noisy vector $\mathbf{r}'_\mathrm{o} = \mathbf{h}_\mathrm{p} * \mathbf{h}_\mathrm{s}$ is compared to the holistic representations of all entities using cosine similarity, according to which the entities are then ranked. In general, multiple objects could be connected to a single subject-predicate pair. Thus, we employ the *filtered mean rank* introduced in [Bordes et al. 2013] to evaluate the memorization task.

We have discussed that the number of pairwise quasi-orthogonal vectors crucially depends on the random initialization. Now we analyse, if the memory capacity depends on the quasi-orthogonality of the initial representation vectors, as well. We perform memorization task on three different KGs, which are FB15k-237 [Toutanova et al. 2015], YAGO3 [Mahdisoltani et al. 2013], and a subset of GDELT [Leetaru et al. 2013]. The exact statistics of the datasets are given in Table. 1.

Table 1: Statistics of KGs

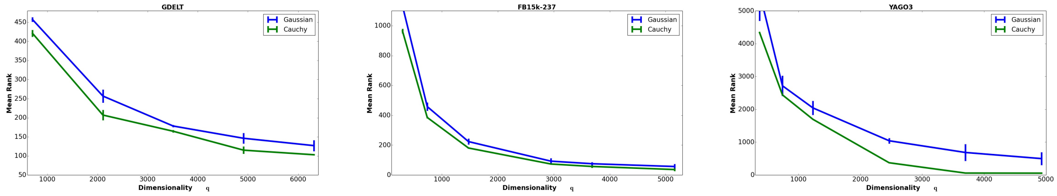
	$\#\mathcal{D}$	N_a	N_e	N_p
GDELT	$497,605$	≈ 73	6786	231
FB15k-237	$301,080$	≈ 20	14505	237
YAGO3	$1,089,000$	≈ 9	123143	37

Recall that N_e and N_p denote the number of entities and predicates, respectively. Moreover, $\#\mathcal{D}$ denotes the total number of triples in a KG, and N_a is the average number of association pairs compressed into holistic feature vectors of entities, which can be estimated as $\frac{\#\mathcal{D}}{N_e}$. After encoding triples in a dataset into holistic features, filtered mean rank is evaluated by ranking retrieved subjects and objects of all triples. Filtered mean ranks on three datasets with holistic representations encoded from Gaussian and Cauchy distributions are displayed in Fig. 3 (a)-(c).

Cauchy holistic representations outperform Gaussian holistic representations significantly when the total number of entities is large (see, Fig. 3(c) for YAGO3), or the average number of encoded associations is large (see, Fig. 3(a) for GDELT). This implies that quasi-orthogonality plays an important role in holographic memory. Improved quasi-orthogonality allows for more entities to be initialized with quasi-orthogonal representations, which is very important for memorizing huge KGs. In addition, it reduces the interference between associations. Moreover, Cauchy holistic features are intrinsically very sparse, making them an attractive candidate for modeling biologically plausible memory systems.

4.4 CORRELATION VERSUS CONVOLUTION

On of the main differences between *holistic representation* and the *holographic reduced representation* is the binding operation. In HRR, two vectors are composed



(a) (b) (c)

Figure 3: Filtered MR vs. the dimensionality of holistic representations evaluated on dataset: (a) GDELT, (b) FB15k-237, and (c) YAGO3. Blues lines denote holistic representations encoded from Gaussian random vectors, and green lines denote holistic representations encoded from Cauchy random vectors. Lower values are preferred.

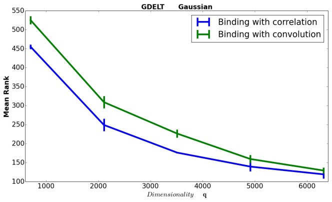


Figure 4: Filtered MR vs. the dimensionality of holistic representations evaluated on the GDELT dataset with Gaussian initialization.

via circular convolution, while in holistic representation, they are composed via circular correlation.

Binding with convolution and correlation is compared in Fig. 4. We report the filtered MR scores on the GDELT dataset versus the dimensionality of holistic representations. It can be seen that binding with circular correlation is significantly superior to convolution. Therefore, a non-commutative compositional operator is essential for storing the directed structures of KG into holographic memory. A theoretical explanation is given in the A.4, along with experimental results on other datasets.

4.5 HYPER-PARAMETER ξ

In the experiments so far, the optimal hyper-parameter ξ is found via grid search. However, it is possible to roughly estimate the range of the optimal hyper-parameter ξ. Indeed, ξ strongly depends on λ_G or λ_C and the average number of encoded association pairs N_a.

So far, the deep relation between holographic memory capacity and quasi-orthogonality has not been discussed in the literature. In the original work on HRR, memory capacity and information retrieval quality are estimated from the distribution of elements in random vectors. In this section we give a plausible explanation from the point of view of the pairwise angle distribution.

Consider a subject s. The predicate-object pair (p, o) is stored in the holistic representation $\mathbf{h}_\mathrm{s}$ along with the other $N_\mathrm{a} - 1$ pairs, such that

$$\mathbf{h}_\mathrm{s} = \xi N_\mathrm{a} \mathbf{r}_\mathrm{s} + \mathbf{r}_\mathrm{p} \star \mathbf{r}_\mathrm{o} + \sum_{i=2}^{N_\mathrm{a}} \mathbf{r}_{p_i} \star \mathbf{r}_{o_i}.$$

Suppose we try to identify the object in the triple $(\mathrm{s}, \mathrm{p}, \cdot)$ via $\mathbf{h}_\mathrm{s}$ and $\mathbf{h}_\mathrm{p}$. After decoding, the noisy vector $\mathbf{r}'_\mathrm{o} = \mathbf{h}_\mathrm{p} * \mathbf{h}_\mathrm{s}$ should be recalled with $\mathbf{h}_\mathrm{o}$, which is the holistic representation of o. Let $\theta_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}}$ denote the angle between $\mathbf{r}'_\mathrm{o}$ and $\mathbf{h}_\mathrm{o}$. The cosine function of this angle is again defined as $\rho_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}} := \cos \theta_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}}$.

In order to recall the object successfully, the angle between $\mathbf{r}'_\mathrm{o}$ and $\mathbf{h_o}$ should be smaller than the expected absolute angle between two arbitrary vectors, namely

$$\theta_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}} < \mathbb{E}[|\theta_{\mathrm{G/C}}|], \tag{13}$$

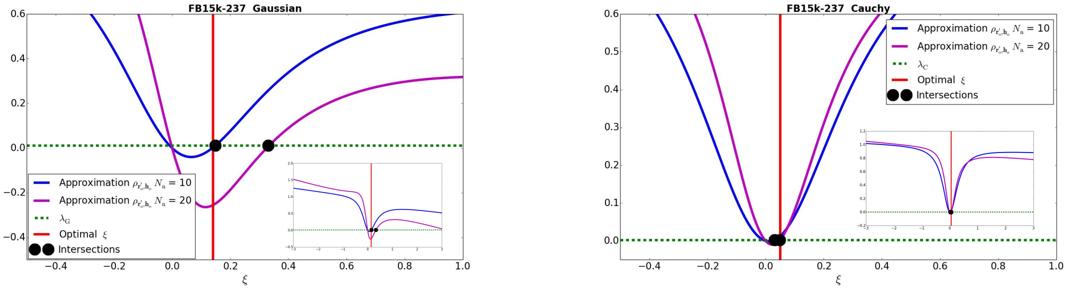
This inequality first implies that the optimal ξ should be a positive number. Given the definition of $\lambda_{\mathrm{G/C}}$ in Eq. 5 and 9, equivalently, Eq. 13 requires

$$\rho_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}} > \lambda_{\mathrm{G/C}}. \tag{14}$$

After some manipulations, a sufficient condition to recognize the object correctly is given by (see A.5)

$$\begin{aligned} &\rho_{\mathbf{r}'_\mathrm{o}, \mathbf{h}_\mathrm{o}} > \\ &\frac{\xi^2 N_\mathrm{a}^2 - (\xi^3 N_\mathrm{a}^3 + 2\xi^2 N_\mathrm{a}^3 - \xi^2 N_\mathrm{a}^2 + \xi N_\mathrm{a}^2 + \xi N_\mathrm{a}^3)\lambda_{\mathrm{G/C}}}{\xi^2 N_\mathrm{a}^2 + N_\mathrm{a} + 2\xi N_\mathrm{a}^2 \lambda_{\mathrm{G/C}} + N_\mathrm{a}(N_\mathrm{a} - 1)\lambda_{\mathrm{G/C}}} \\ &> \lambda_{\mathrm{G/C}}. \end{aligned} \tag{15}$$

In the following, we verify this condition on the FB15k-237 dataset. We consider one of the experimental settings employed in the memorization task. The dimension of holistic features is $q = 5200$, with $\lambda_\mathrm{G} = 0.0111$ computed from Eq. 5, and $\lambda_\mathrm{C} = 0.00204$ from Eq. 9. For Gaussian initialization, the optimum is found at $\xi = 0.14$ via grid search, while for Cauchy initialization, the optimum is found at $\xi = 0.05$.



(a)

(b)

Figure 5: Analysis of the hyper-parameter ξ on the FB15k-237 dataset. (a): Approximation of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}$ for Gaussian initialization. Curves with $N_a = 10$ (blue), $N_a = 20$ (magenta) and their intersections with the retrieval threshold λ_G are displayed. The red vertical line denotes the experimentally determined optimal ξ. Insert shows the curves with $\xi \in [-3, 3]$. (b): Approximation of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}$ for Cauchy initialization with $N_a = 10$ (blue), and $N_a = 20$ (magenta). Rest remains the same.

To verify these optima, Fig. 5 (a) and (b) display the approximation of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a)$ as a function of ξ. [3] Its intersection with $\lambda_{G/C}$ is marked with a black dot. In FB15k-237, N_a is estimated to be 20, while, in general, a KG could be quite imbalanced. Thus, $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a)$ with $N_a = 10$, and 20 are shown together for comparison.

In Fig. 5 (a) for Gaussian initialization, experimentally determined optimal ξ (red vertical line) is found close to the intersection of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a = 10)$ and threshold λ_G, meaning that Gaussian holistic features tend to memorize fewer association pairs. They can only map sparsely connected graph structures into meaningful representations.

In Fig. 5 (b) for Cauchy initialization, however, the optimal ξ is close to the intersection of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a = 20)$ and λ_C. Thus, Cauchy holistic features are more suitable to memorize a larger chunk of associations, meaning that they are capable of mapping densely connected graph structures into meaningful representations. All optima are found near the intersection points instead of the local maximum with $\xi > 0$. It indicates that, to maximize the memory capacity, the holistic features can only store information with very low fidelity.

Table 2: Filtered recall scores on FB15k-237

			Hits		
Methods	MR	MRR	@10	@3	@1
RESCAL	996	0.221	0.363	0.237	0.156
DISTMULT	254	0.241	0.419	0.263	0.155
COMPLEX	339	0.247	0.428	0.275	0.158
R-GCN [4]	-	0.248	0.414	0.258	0.153
HOLNN$_G$ [5]	235	0.285	0.455	0.315	0.207
HOLNN$_C$	**228**	**0.295**	**0.465**	**0.320**	**0.212**

5 INFERENCE ON KG

5.1 INFERENCE VIA HOLISTIC REPRESENTATION

In this section, we describe the model for inferring the missing links in the KG. Recall the scoring function η_{spo} defined in Sec. 2. Our model uses holistic representations as input and generalizes them to implicit facts, by a two-layer neural network [6]. Formally, the scoring function is given as follow:

$$\begin{aligned}\eta_{spo} = &\langle \mathrm{ReLU}(\mathbf{h}_s \mathbf{W}_1^e)\mathbf{W}_2^e,\ \mathrm{ReLU}(\mathbf{h}_p \mathbf{W}_1^p)\mathbf{W}_2^p, \\ &\mathrm{ReLU}(\mathbf{h}_o \mathbf{W}_1^e)\mathbf{W}_2^e\rangle, \end{aligned} \tag{16}$$

where $\langle \cdot, \cdot, \cdot \rangle$ denotes tri-linear dot product; $\mathbf{h}_s$, $\mathbf{h}_o$ are the holistic representations for entities defined in Eq. 11, $\mathbf{h}_p$ is defined in Eq. 12.

Suppose that the holistic representations are defined in $\mathbb{R}^q$. Then $\mathbf{W}_1^e \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^e \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for entities; $\mathbf{W}_1^p \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^p \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for predicates. We refer Eq. 16 as HOLNN, a combination of holistic representations and a simple neural network.

As an example, for training on FB15k-237, we take $q = 3600$, $h_1 = 64$, and $h_2 = 256$. Note that only weight matrices in the neural network are trainable parameters, holistic representations are fixed after encoding. Thus, the total number of trainable parameters in HOLNN is $0.48M$, which is much smaller than COM-

[3] The approximation of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}$ is the second term of Eq. 15

[4] see [Schlichtkrull et al. 2018]

[5] G stands for Gaussian holistic features, and C for Cauchy holistic features.

[6] Further experimental details are referred to A.8

plEx with $5.9M$ parameters, by assuming that the dimension of embeddings in the ComplEx is 200.

To evaluate the performance of HolNN for missing links prediction, we compare it to the state-of-the-art models on two datasets: FB15k-237, and GDELT. They were split randomly in training, validation, and test sets. We implement all models with the identical loss function Eq. 1, and minimize the loss on the training set using Adam as the optimization method. Hyper-parameters, e.g., the learning rate, and $l2$ regularization, are optimized based on the validation set.

We use filtered MR, filtered mean reciprocal rank (MRR), and filtered Hits at n (Hits@n) as evaluation metrics [Bordes et al. 2013]. Table 2 and Table 3 report different metrics on the FB15k-237, and GDELT dataset, respectively. It can be seen that HolNN is superior to all the baseline methods on both datasets with considerably less trainable parameters. Moreover, HolNN_{C} consistently outperforms HolNN_{G}, indicating that the memory capacity of holistic representations is important for generalization.

Table 3: Filtered recall scores on GDELT

Methods	MR	MRR	Hits		
			@10	@3	@1
RESCAL	212	0.202	0.396	0.225	0.107
DistMult	181	0.232	0.451	0.268	0.124
ComplEx	158	0.256	0.460	0.295	0.146
HolNN_{G}	105	0.284	0.457	0.301	0.198
HolNN_{C}	**102**	**0.296**	**0.471**	**0.315**	**0.210**

5.2 INFERENCE ON NEW ENTITIES

In additional experiments, we show that HolNN is capable of inferring implicit facts on new entities without retraining the neural network. Experiments are performed on FB15k-237 as follows. We split the entire FB15k-237 dataset $\mathcal{D}$ into $\mathcal{D}_{\text{old}}$ and $\mathcal{D}_{\text{new}}$. In $\mathcal{D}_{\text{new}}$, the subjects of triples are new entities which do not show up in $\mathcal{D}_{\text{old}}$, while objects and predicates are already seen in the $\mathcal{D}_{\text{old}}$. Suppose our task is to predict implicit links between new entities (subjects in $\mathcal{D}_{\text{new}}$) and old entities (entities in $\mathcal{D}_{\text{old}}$). Thus, we further split $\mathcal{D}_{\text{new}}$ into $\mathcal{D}_{\text{new}}^{\text{train}}$, $\mathcal{D}_{\text{new}}^{\text{valid}}$, and $\mathcal{D}_{\text{new}}^{\text{test}}$ sets.

For embedding models, e.g., ComplEx, after training on $\mathcal{D}_{\text{old}}$, the most efficient way to solve this task is to adapt the embeddings of new entities on $\mathcal{D}_{\text{new}}^{\text{train}}$, with fixed embeddings of old entities. On the other hand, for the HolNN model, new entities obtain their holistic representations via triples in the $\mathcal{D}_{\text{new}}^{\text{train}}$ set. These holistic features are then fed into the trained two-layer neural network. Table 4 shows filtered recall scores for predicting links between new entities and old entities on $\mathcal{D}_{\text{new}}^{\text{test}}$, with the number of new entities in $\mathcal{D}_{\text{new}}$ being 300, 600, or 900. ComplEx and HolNN with Cauchy holistic features are compared.

There are two settings for the HolNN_{C} model. New entities could be encoded either from holistic features of old entities, or from random initializations of old entities [7]. We denote these two cases as $\text{HolNN}_{\text{C}}(\mathbf{h})$ and $\text{HolNN}_{\text{C}}(\mathbf{r})$, respectively. It can be seen that $\text{HolNN}_{\text{C}}(\mathbf{r})$ outperforms $\text{HolNN}_{\text{C}}(\mathbf{h})$ only to some degree. It indicates that HolNN_{C} is robust to the noise, making it generalizes well.

Table 4: Inference of new entities on FB15k-237

	Number of New Entities					
	300		600		900	
Methods	MR	MRR	MR	MRR	MR	MRR
ComplEx	262	0.291	**265**	0.266	**286**	0.243
$\text{HolNN}_{\text{C}}(\mathbf{h})$	345	0.274	415	0.242	510	0.222
$\text{HolNN}_{\text{C}}(\mathbf{r})$	**252**	**0.315**	302	**0.281**	395	**0.265**

6 CONCLUSION

We have introduces the holistic representation for the distributed storage of complex association patterns and have applied it to knowledge graphs. We have shown that interference between stored information is reduced with initial random vectors which are pairwise quasi-orthogonal and that pairwise quasi-orthogonality can be improved by drawing vectors from heavy-tailed distributions, e.g., a Cauchy distribution. The experiments demonstrated excellent performance on memory retrieval and competitive results on link prediction.

In our approach, latent representations are derived from random vectors and are not learned from data, as in most modern approaches to representation learning on knowledge graphs. One might consider representations derived from random vectors to be biologically more plausible, if compared to representations which are learned via complex gradient based update rules. Thus in addition to its very competitive technical performance, one of the interesting aspects of our approach is its biological plausibility.

Outlook: Potential applications could be applying the holistic encoding algorithm to Lexical Functional for modeling distributional semantics [Coecke et al. 2010], or graph convolutional network [Kipf et al. 2017] for semi-supervised learning using holistic representations as feature vectors of nodes on a graph.

[7] Recall that random initializations are actually deleted after encoding. Here we use them just for comparison.

References

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor (2008). "Freebase: a collaboratively created graph database for structuring human knowledge". *Proceedings of the 208 ACM SIGMOD*. AcM, pp. 1247–1250.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko (2013). "Translating embeddings for modeling multi-relational data". *NIPS*, pp. 2787–2795.

Cai, Tony and Tiefeng Jiang (2012). "Phase transition in limiting distributions of coherence of high-dimensional random matrices". *Journal of Multivariate Analysis* 107, pp. 24–39.

Cai, Tony, Jianqing Fan, and Tiefeng Jiang (2013). "Distributions of angles in random packing on spheres". *The Journal of Machine Learning Research* 14.1, pp. 1837–1864.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark (2010). "Mathematical foundations for a compositional distributional model of meaning". *Linguistic Analysis 36*.

Diaconis, Persi and David Freedman (1984). "Asymptotics of graphical projection pursuit". *The annals of statistics*, pp. 793–815.

Gabor, D. (1969). "Associative holographic memories". *IBM Journal of Research and Development* 13.2, pp. 156–159.

Gentner, Dedre (1983). "Structure-mapping: A theoretical framework for analogy". *Cognitive science* 7.2, pp. 155–170.

Gnedenko, B.V. and A.N. Kolmogorov (1954). *Limit distributions for sums of independent random variables*. Addison-Wesley.

Hall, Peter, James Stephen Marron, and Amnon Neeman (2005). "Geometric representation of high dimension, low sample size data". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.3, pp. 427–444.

Hopfield, John J. (1982). "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the national academy of sciences* 79.8, pp. 2554–2558.

Kipf, Thomas N and Max Welling (2017). "Semi-supervised classification with graph convolutional networks". *ICLR*.

Leetaru, Kalev and Philip A. Schrodt (2013). "GDELT: Global data on events, location, and tone". *ISA Annual Convention*.

Ma, Yunpu, Volker Tresp, and Erik Daxberger (2018). "Embedding models for episodic memory". *arXiv preprint arXiv:1807.00228*.

Mahdisoltani, Farzaneh, Joanna Biega, and Fabian M. Suchanek (2013). "Yago3: A knowledge base from multilingual wikipedias". *CIDR*.

McClelland, James L., Bruce L. McNaughton, and Randall C. O'reilly (1995). "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." *Psychological review* 102.3, p. 419.

Muirhead, Robb J. (2009). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel (2011). "A Three-Way Model for Collective Learning on Multi-Relational Data". *ICML*. Vol. 11, pp. 809–816.

Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel (2012). "Factorizing yago: scalable machine learning for linked data". *Proceedings of the 21st international conference on World Wide Web*. ACM, pp. 271–280.

Nickel, Maximilian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich (2016a). "A review of relational machine learning for knowledge graphs". *Proceedings of the IEEE* 104.1, pp. 11–33.

Nickel, Maximilian, Lorenzo Rosasco, and Tomaso Poggio (2016b). "Holographic Embeddings of Knowledge Graphs". *AAAI*, pp. 1955–1961.

Penrose, Mathew (2003). *Random geometric graphs*. 5. Oxford university press.

Plate, Tony A. (1995). "Holographic reduced representations". *IEEE Transactions on Neural networks* 6.3, pp. 623–641.

Schlichtkrull, Michael, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling (2018). "Modeling relational data with graph convolutional networks". *European Semantic Web Conference*. Springer, pp. 593–607.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (2007). "Yago: a core of semantic knowledge". *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706.

Toutanova, Kristina and Danqi Chen (2015). "Observed versus latent features for knowledge base and text inference". *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66.

Tresp, Volker, Yunpu Ma, Stephan Baier, and Yinchong Yang (2017a). "Embedding learning for declarative memories". *ESWC*. Springer, pp. 202–216.

Tresp, Volker and Yunpu Ma (2017b). "The Tensor Memory Hypothesis". *arXiv preprint arXiv:1708.02918*.

Trouillon, Théo, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard (2016). "Complex embeddings for simple link prediction”. *ICML*, pp. 2071–2080.

Westlake, Philip R. (1970). “The possibilities of neural holographic processes within the brain”. *Kybernetik* 7.4, pp. 129–153.

Yang, Bishan, Wentau Yih, Xiaodong He, Jianfeng Gao, and Li Deng (2014). “Embedding entities and relations for learning and inference in knowledge bases”. *ICLR 2015*.

A APPENDIX

A.1 DERIVATION OF COROLLARY 1 & 2

Corollary 1. *Consider a set of independent q-dimensional Gaussian random vectors which are pairwise ϵ-orthogonal with probability $1-\nu$, then the number of such Gaussian random vectors is bounded by*

$$N \le \sqrt[4]{\frac{\pi}{2q}}\, \mathrm{e}^{\frac{\epsilon^2 q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \qquad \text{(A.1)}$$

Proof. Recall that, in the case of Gaussian distributed random vectors, the pdf of ρ is

$$g(\rho) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} (1-\rho^2)^{\frac{q-3}{2}}.$$

This directly yields that $\omega := \sqrt{q}\rho$ has the density function

$$f(\omega) = \frac{1}{\sqrt{q}} \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} \left(1 - \frac{\omega^2}{q}\right)^{\frac{q-3}{2}} \to \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\omega^2}{2}} \qquad \text{(A.2)}$$

as $q \to \infty$, using the fact that $\frac{\Gamma(\frac{q}{2})}{\Gamma(\frac{q-1}{2})} \sim \sqrt{\frac{q}{2}}$. Therefore the probability that two random Gaussian vectors are not ϵ-orthogonal is upper bounded by

$$\begin{aligned} \Pr(|\rho| \ge \epsilon) &= \Pr(|\omega| \ge \sqrt{q}\epsilon) = 2 \int_{\sqrt{q}\epsilon}^{\sqrt{q}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\omega^2}{2}}\, \mathrm{d}\omega \\ &< \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{q\epsilon^2}{2}} (\sqrt{q} - \sqrt{q}\epsilon) < \sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{q\epsilon^2}{2}}. \end{aligned} \qquad \text{(A.3)}$$

To estimate the probability that ϵ-orthogonality is satisfied for a set of N independent Gaussian random vectors, let us consider the following quantity

$$\mathcal{P}(\epsilon, N) := \prod_{k=1}^{N-1} \left[1 - k \Pr(|\rho| \ge \epsilon)\right]. \qquad \text{(A.4)}$$

The above estimation has clear meaning. Given one Gaussian random vector $\mathbf{X}_1$, the probability that an independently sampled random vector $\mathbf{X}_2$ which is not ϵ-orthogonal to $\mathbf{X}_1$ is $\Pr(|\rho| > \epsilon)$. Similarly, given k i.i.d. Gaussian random vectors $\mathbf{X}_1, \cdots, \mathbf{X}_k$, the probability that an independently drawn Gaussian random vector $\mathbf{X}_{k+1}$ which is not ϵ-orthogonal to $\mathbf{X}_1, \cdots, \mathbf{X}_k$ is upper bounded by $k \Pr(|\rho| > \epsilon)$. Therefore, we have the estimate in Eq. A.4 for N independent random vectors. Using Eq. A.3, $\mathcal{P}(\epsilon, N)$ can be computed as follows

$$\begin{aligned} \mathcal{P}(\epsilon, N) &> \prod_{k=1}^{N-1} (1 - k\sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}}) \\ &> (1 - N\sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}})^N \sim \mathrm{e}^{-N^2 \sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}}}, \end{aligned}$$

for sufficiently large N and q satisfying $N\sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}} < 1$. If we require $\mathcal{P}(\epsilon, N) \ge 1 - \nu$, then the number of pairwise ϵ-orthogonal i.i.d. Gaussian random vectors is bounded from above by

$$\begin{aligned} &\mathrm{e}^{-N^2 \sqrt{\frac{2q}{\pi}} \mathrm{e}^{-\frac{\epsilon^2 q}{2}}} \ge 1 - \nu \quad \Rightarrow \\ &N \le \sqrt[4]{\frac{\pi}{2q}}\, \mathrm{e}^{\frac{\epsilon^2 q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}} \end{aligned}$$

■

Corollary 2. *Consider a set of n q-dimensional random Gaussian vectors, we have*

$$\lambda_{\mathrm{G}} := \mathbb{E}[|\rho_{\mathrm{G}}|] = \sqrt{\frac{2}{\pi q}}. \qquad \text{(A.5)}$$

Proof. Given the $g(\rho_{\mathrm{G}})$ in Theorem 1, we have

$$\begin{aligned} \mathbb{E}[|\rho_{\mathrm{G}}|] &= \int_{-1}^{1} |\rho| g(\rho)\, \mathrm{d}\rho = \sqrt{\frac{2q}{\pi}} \int_0^1 \rho(1-\rho^2)^{\frac{q-3}{2}}\, \mathrm{d}\rho \\ &= -\sqrt{\frac{2q}{\pi}} \frac{(1-\rho^2)^{\frac{q-1}{2}}}{q-1} \Bigg|_0^1 = \sqrt{\frac{2}{\pi q}}, \end{aligned}$$

for large q. ■

A.2 DISCUSSION ON CONJECTURE 1

In this section, we derive the approximations stated in Conjecture 1 and verify them with empirical simulations.

According to the central limit theorem, the sum of independently and identically distributed random variables with finite variance converges weakly to a normal distribution as the number of random variables approaches infinity. Our derivation relies on the generalized central limit theorem proven by Gnedenko and Kolmogorov in 1954 [Gnedenko et al. 1954].

Theorem A 1. *(**Generalized Central Limit Theorem** [Gnedenko et al. 1954]) Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d random variables drawn from the distribution with probability density function $f(x)$ with the following asymptotic behaviour*

$$f(x) \simeq \begin{cases} c_+ x^{-(\alpha+1)} & \textit{for} \quad x \to \infty \\ c_- |x|^{-(\alpha+1)} & \textit{for} \quad x \to -\infty, \end{cases} \qquad \text{(A.6)}$$

where $0 < \alpha < 2$, and c_+, c_- are real positive numbers. Define random variable S_n as a superposition of $X_1, \cdots, X_n$

$$S_n = \frac{\sum\limits_{i=1}^{n} X_i - C_n}{n^{\frac{1}{\alpha}}}, \quad \textit{with}$$

$$C_n = \begin{cases} 0 & \textit{if} \quad 0 < \alpha < 1 \\ n^2 \Im \ln(\phi_X(1/n)) & \textit{if} \quad \alpha = 1 \\ n\mathbb{E}[X] & \textit{if} \quad 1 < \alpha < 2, \end{cases}$$

where ϕ_X is the characteristic function of a random variable X with probability density function $f(x)$, $\mathbb{E}[X]$ is the expectation value of X, $\Im$ denotes the imaginary part of a variable. Then as the number of summands n approaches infinity, the random variables S_n converge in distribution to a unique stable distribution $S(x; \alpha, \beta, \gamma, 0)$, that is

$$S_n \xrightarrow{d} S(\alpha, \beta, \gamma, 0), \quad \textit{for} \quad n \to \infty,$$

where, α characterizes the power-law tail of $f(x)$ as defined above, and parameters β and γ are given as:

$$\begin{aligned} \beta &= \frac{c_+ - c_-}{c_+ + c_-}, \\ \gamma &= \left[\frac{\pi(c_+ + c_-)}{2\alpha \sin(\frac{\pi\alpha}{2})\Gamma(\alpha)} \right]^{\frac{1}{\alpha}}. \end{aligned} \tag{A.7}$$

To be self-contained, we give the definition of stable distributions after [Nolan 2003; Mandelbrot 1960].

Definition A 1. *A random variable X follows a stable distribution if its characteristic function can be expressed as*

$$\phi(t; \alpha, \beta, \gamma, \mu) = \mathrm{e}^{i\mu t - |\gamma t|^{\alpha}(1 - i\beta \ \mathrm{sgn}(t)\Phi(\alpha, t))}, \tag{A.8}$$

with $\Phi(\alpha, t)$ defined as

$$\Phi(\alpha, t) = \begin{cases} \tan(\frac{\pi\alpha}{2}) & \textit{if} \quad \alpha \neq 1 \\ -\frac{2}{\pi} \log |t| & \textit{if} \quad \alpha = 1. \end{cases}$$

Then the probability density function $S(x; \alpha, \beta, \gamma, \mu)$ of the random variable X is given by the Fourier transform of its characteristic function

$$S(x; \alpha, \beta, \gamma, \mu) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \phi(t; \alpha, \beta, \gamma, \mu) \, \mathrm{e}^{-ixt} \, \mathrm{d}x.$$

The parameter α satisfying $0 < \alpha \leq 2$ characterizes the power-law asymptotic limit of the stable distribution, $\beta \in [-1, 1]$ measures the skewness, $\gamma > 0$ is the scale parameter, and $\mu \in \mathbb{R}$ is the shift parameter. Note that the normal distribution is a typical stable distribution. Other examples with analytical expression include the Cauchy distribution and the Lévy distribution. For the later use, we give the analytical form of the Lévy distribution.

Remark A 1. *The probability density function of the Lévy distribution is given by*

$$f(x; \gamma, \mu) = \sqrt{\frac{\gamma}{2\pi}} \frac{\mathrm{e}^{-\frac{\gamma}{2(x-\mu)}}}{(x - \mu)^{\frac{3}{2}}}, \quad x \geq \mu, \tag{A.9}$$

where μ is the shift parameter and γ is the scale parameter. The Lévy distribution is a special case of the stable distribution $S(x; \alpha, \beta, \gamma, \mu)$ with $\alpha = \frac{1}{2}$ and $\beta = 1$. This can be seen from its characteristic function, which can be written as

$$\phi(t; \gamma, \mu) = \mathrm{e}^{i\mu t - |\gamma t|^{1/2}(1 - i \ \mathrm{sgn}(t))}$$

To derive $g(\rho_{\mathrm{C}})$ for Cauchy random vectors, we first need the distribution function of X^2 given that the random variable X has a Cauchy distribution.

Lemma A 1. *Let X be a Cauchy random variable having the probability density function $f_X(x) = \frac{1}{\pi} \frac{\zeta}{x^2 + \zeta^2}$, where $\zeta > 0$ is the scale parameter. Then the squared variable $Y := X^2$ has the pdf:*

$$f_Y(y) = \begin{cases} \frac{1}{\pi} \frac{\zeta}{\sqrt{y}(\zeta^2 + y)} & \textit{for} \quad y \geq 0, \\ 0 & \textit{otherwise.} \end{cases} \tag{A.10}$$

Proof. $f_Y(y)$ can be derived from $f_X(x)$ by a simple variable transformation $y = g(x) = x^2$. In particular, utilizing the symmetry of $f_X(x)$, we have

$$\begin{aligned} f_Y(y) &= 2 \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| f_X(g^{-1}(y)) \\ &= \frac{1}{\pi} \frac{\zeta}{\sqrt{y}(\zeta^2 + y)}. \end{aligned}$$

■

In the following Lemma we derive the probability density function for $z_{\mathbf{X},\mathbf{Y}}$, which is defined as $z_{\mathbf{X},\mathbf{Y}} := \frac{1}{q^2} \frac{X_2^2 + \cdots X_q^2}{X_1^2}$.

Lemma A 2. *Let $X_1, \cdots, X_q$ be a sequence of i.i.d. random variables drawn from $\mathcal{C}(0, 1)$. Then the random variable $Z_q := \frac{1}{q^2} \frac{X_2^2 + \cdots + X_q^2}{X_1^2}$ converges in distribution to*

$$f(z) = -\frac{1}{\pi^2} \frac{1}{z^{\frac{3}{2}}} \left[\mathrm{e}^{\frac{1}{\pi z}} \ \mathrm{Ei}\left(-\frac{1}{\pi z} \right) \right], \tag{A.11}$$

as $q \to \infty$, where $\mathrm{Ei}(x)$ denotes the exponential integral.

Proof. The numerator in Z_q can be regarded as a sum of independent random variables with density function $f_{Y:=X^2}(y) = \frac{1}{\pi}\frac{1}{\sqrt{y}(1+y)}$, see Eq. A.10 with $\zeta = 1$. Thus, we can use the generalized central limit theorem to obtain the density function $g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2)$ for the numerator, as $q \to \infty$.

Note that $f_Y(y) \sim \frac{1}{\pi}y^{-\frac{3}{2}}$ as $y \to +\infty$. From this asymptotic behaviour we can extract that $c_+ = \frac{1}{\pi}$, $c_- = 0$, and $\alpha = \frac{1}{2}$. Moreover, Eq. A.7 with $\beta = 1$ yields $\gamma = \left[\frac{1}{\sin(\frac{\pi}{4})\,\Gamma(\frac{1}{2})}\right]^2 = \frac{2}{\pi}$. In summary, $g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2)$ converges to a unique stable distribution $S(\alpha = \frac{1}{2}, \beta = 1, \gamma = \frac{2}{\pi}, \mu = 0)$, which is exactly the Lévy distribution shown in Remark A 1. Hence, we have

$$g(\frac{1}{q^2}\sum_{i=2}^{q} X_i^2) \xrightarrow{d} S(x; \frac{1}{2}, 1, \frac{2}{\pi}, 0) = \frac{1}{\pi}\frac{\mathrm{e}^{-\frac{1}{\pi x}}}{x^{\frac{3}{2}}}, \qquad \text{as} \quad q \to \infty. \tag{A.12}$$

Next, we consider the quotient distribution of two random variables in order to derive the pdf of Z_q. To be more specific, let X and Y be independent non-negative random variables with corresponding probability density function $f_X(x)$ and $f_Y(y)$ over the domains $x \geq 0$ and $y \geq 0$, respectively. Then the cumulative distribution function $F_Z(z)$ of $Z := \frac{Y}{X}$ can be computed by

$$\begin{aligned} F_Z(z) &= \Pr(\frac{Y}{X} \leq z) = \Pr(Y \leq zX) \\ &= \int_0^\infty \left[\int_0^{y=zx} f_Y(y)\mathrm{d}y\right] f_X(x)\mathrm{d}x. \end{aligned}$$

Differentiating the cumulative distribution function yields

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) = \int_0^\infty x\, f_Y(zx)\, f_X(x)\, \mathrm{d}x.$$

Following the above procedure, we can obtain the pdf for Z_q as $q \to \infty$ in case the density functions of the numerator and the denominator are given by Eq. A.12 and Eq. A.10, respectively. That yields

$$\begin{aligned} f(z) &= \frac{1}{\pi^2}\int_0^\infty x\,\frac{\mathrm{e}^{-\frac{1}{\pi z x}}}{(zx)^{\frac{3}{2}}}\,\frac{1}{\sqrt{x}(1+x)}\,\mathrm{d}x \\ &= \frac{1}{\pi^2}\frac{1}{z^{\frac{3}{2}}}\left[-\mathrm{e}^{\frac{1}{\pi z}}\,\mathrm{Ei}\left(-\frac{x+1}{\pi z x}\right)\right]\Bigg|_{x=0}^{\infty} \\ &= -\frac{1}{\pi^2}\frac{1}{z^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z}}\,\mathrm{Ei}\left(-\frac{1}{\pi z}\right)\right]. \end{aligned}$$

■

In the following we discuss why the density function $g(\rho_{\mathrm{C}})$ can only be approximated by taking the limit as $q \to \infty$.

Suppose $\mathbf{X} = (X_1, \cdots, X_q)$ and $\mathbf{Y} = (Y_1, \cdots, Y_q)$ are Gaussian random variables. To derive $g(\rho_{\mathbf{X},\mathbf{Y}})$ in Lemma 1, [Cai et al. 2012; Muirhead 2009] compute the density function of $\frac{\boldsymbol{\alpha}^{\mathsf{T}}\cdot\mathbf{X}}{||\mathbf{X}||}$ instead, where $\boldsymbol{\alpha}^{\mathsf{T}}\cdot\boldsymbol{\alpha} = 1$, and $\boldsymbol{\alpha} := \frac{\mathbf{Y}}{||\mathbf{Y}||}$. In particular, without loss of generality, they assume $\boldsymbol{\alpha} = (1, 0, \cdots, 0)$. The justification for this assumption is that the random variable $\mathbf{X}' := \frac{\mathbf{X}}{||\mathbf{X}||}$ is uniformly distributed on the $(q-1)$-dimensional sphere (see Theorem 1.5.6 in [Muirhead 2009]).

In our case, the distributional uniformity of $\frac{\mathbf{X}}{||\mathbf{X}||}$ is not superficial, since the density function of $\mathbf{X}'$ doesn't depend on $\mathbf{X}'$ only through the value of $\mathbf{X}'^{\mathsf{T}}\mathbf{X}'$. To see this, in the following Lemma, we discuss the distribution function of the normalization $\frac{\mathbf{X}}{||\mathbf{X}||}$.

Lemma A 3. *Consider a q-dimensional random vector $\mathbf{X} = (X_1, \cdots, X_q)$, where $X_1, \cdots, X_q$ are independently and identically drawn from a Cauchy distribution $\mathcal{C}(0, 1)$. Then, as $q \to \infty$, the normalized random vector $\frac{\mathbf{X}}{||\mathbf{X}||} = (X_1', \cdots, X_q')$ has a joint density function, in which the random variables $X_1', \cdots, X_q'$ are all independent from each other.*

Proof. Without loss of generality, we study the pdf of $X_1' = \frac{X_1}{\sqrt{X_1^2 + \cdots + X_q^2}}$. Similar to the proof of Lemma A 2, the random variable $Z_q := \frac{1}{q^2}\frac{X_2^2 + \cdots + X_q^2}{X_1^2}$ converges weakly to the distribution with pdf given by Eq. A.11 as $q \to \infty$, which is independent of the other random variables due to the generalized central limit theorem. Hence, X_1' can be treated as an independent random variable as $q \to \infty$. In addition, we obtain the pdf of X_1' given by

$$f_{X_1'}(x_1') = -\frac{2}{\pi^2 q^2 x_1'^3}\frac{1}{z_1^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z_1}}\,\mathrm{Ei}\left(-\frac{1}{\pi z_1}\right)\right], \tag{A.13}$$

where z_1 is defined as $z_1 := \frac{1}{q^2}\left(\frac{1}{x_1'^2} - 1\right)$. The arguments can be easily generalized to $X_2', \cdots, X_q'$. ■

The pdf of the joint distribution $f_{\mathbf{X}'}(x_1', \cdots, x_q')$ can be written as a product of marginals, that is

$$f_{\mathbf{X}'}(x_1', \cdots, x_q') = \prod_{i=1}^{q} f_{X_i'}(x_i'),$$

as $q \to \infty$. The density function of $\mathbf{X}'$ is not invariant under an arbitrary rotation. Thus, it is not uniformly distributed on S^{q-1}.

The above density function of normalized Cauchy random vectors leads to the following Remark.

Remark A 2. *The normalized Cauchy random vector $\mathbf{X}' = \frac{\mathbf{X}}{||\mathbf{X}||}$ is sparse in the sense that the density function of its elements can be approximated by a δ-function.*

Fig. 1 shows the empirical elements distribution of 1000 normalized Cauchy random vectors. This indicates that in sufficiently high-dimensional spaces the density function of the normalized entries converges to a δ-function. To explain this, recall the Laurent expansion of the density function given in Eq. A.13,

$$f_{X_1'}(x_1') = \frac{2}{\pi q x_1'^2} - \frac{2}{q^3 x_1'^4} + \frac{4\pi}{q^5 x_1'^6} + \mathcal{O}\left(\frac{1}{q^7 x_1'^8}\right). \tag{A.14}$$

This expansion converges to zero almost everywhere expect for $x_1' = 0$ as $q \to \infty$.

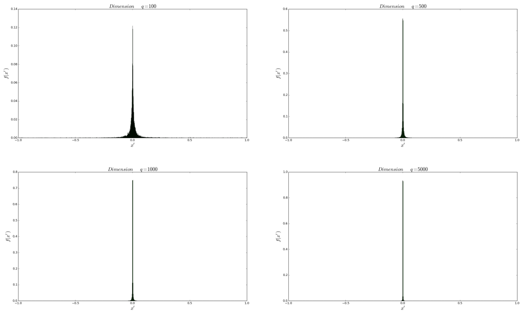


Figure 1: Empirical distributions of 10000 normalized Cauchy random vectors with dimensions q = 100, 500, 1000, 5000.

In the following, we provide a full derivation of $g(\rho_\mathrm{C})$ proposed in the Conjecture 1.

Conjecture 1. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent q-dimensional random vectors whose elements are independently and identically drawn from a Cauchy distribution $\mathcal{C}(0,1)$. Let Θ_{ij} be the angle between $\mathbf{X}_i$ and $\mathbf{X}_j$. Then, as $q \to \infty$, $\rho_{ij} := \cos\Theta_{ij} \in [-1,1]$, $1 \leq i < j \leq n$ are pairwise i.i.d. with density function approximated by*

$$g(\rho_\mathrm{C}) = -\frac{2}{\pi^2 q^2 \rho_\mathrm{C}^3} \cdot \frac{1}{z^{\frac{3}{2}}}\left[\mathrm{e}^{\frac{1}{\pi z}}\,\mathrm{Ei}\left(-\frac{1}{\pi z}\right)\right], \tag{A.15}$$

where $z := \frac{1}{q^2}\left(\frac{1}{\rho_\mathrm{C}^2} - 1\right)$.

Given two Cauchy random vectors $\mathbf{X} = (X_1, \cdots, X_q)$ and $\mathbf{Y} = (Y_1, \cdots, Y_q)$, $\rho_{\mathbf{X},\mathbf{Y}}$ is approximated by $\rho_{\mathbf{X},\mathbf{Y}} \approx \frac{X_1}{\sqrt{X_1^2 \cdots + X_q^2}}$.

Furthermore, we introduce the new variable $z_{\mathbf{X},\mathbf{Y}} := \frac{1}{q^2}(\frac{1}{\rho_{\mathbf{X},\mathbf{Y}}} - 1)$. From Lemma A 2 we have the density function $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$. Then, $g(\rho_{\mathbf{X},\mathbf{Y}})$ can be directly obtained from $\hat{g}(z_{\mathbf{X},\mathbf{Y}})$ by a variable transform, that is $g(\rho_{\mathbf{X},\mathbf{Y}}) = \left|\frac{\mathrm{d}z}{\mathrm{d}\rho}\right| \hat{g}(z_{\mathbf{X},\mathbf{Y}})$. With $\left|\frac{\mathrm{d}z}{\mathrm{d}\rho}\right| = \frac{2}{q^2\rho^3}$ we immediately get Eq. A.15 as the density function for $\rho_{\mathbf{X},\mathbf{Y}}$.

Assume that Eq. A.15 is valid as $q \to \infty$. In the following we show that $\{\rho_{ij} | 1 \leq i < j \leq n\}$ are i.i.d random variables. First, notice that ρ_{ij} and ρ_{kl} are independent if $\{i,j\} \cap \{k,l\} = \emptyset$. It is left to prove that $\rho_{\mathbf{X},\mathbf{Y}}$ and $\rho_{\mathbf{X},\mathbf{Z}}$ are independent, given that $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are independent random variables.

To prove the independence, consider $\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})h_2(\rho_{\mathbf{X},\mathbf{Z}})]$, where h_1 and h_2 are arbitrary bounded functions. Since $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are independent,

$$\begin{aligned}\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot & h_2(\rho_{\mathbf{X},\mathbf{Z}})] \\ &= \mathbb{E}\left[\,\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot h_2(\rho_{\mathbf{X},\mathbf{Z}}) | \mathbf{X}]\,\right] \\ &= \mathbb{E}\left[\,\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) | \mathbf{X}] \cdot \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}}) | \mathbf{X}]\,\right].\end{aligned}$$

Given $\mathbf{X}$, the probability density function of $\rho_{\mathbf{X},\mathbf{Y}}$ is independent of $\mathbf{X}$. Thus, $\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})|\mathbf{X}] = \int_{-1}^{1} h_1(\rho_{\mathbf{X},\mathbf{Y}}) g(\rho_{\mathbf{X},\mathbf{Y}})\ \mathrm{d}\rho = \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})]$, and similarly $\mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})|\mathbf{X}] = \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})]$. It gives,

$$\mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}}) \cdot h_2(\rho_{\mathbf{X},\mathbf{Z}})] = \mathbb{E}[h_1(\rho_{\mathbf{X},\mathbf{Y}})] \cdot \mathbb{E}[h_2(\rho_{\mathbf{X},\mathbf{Z}})],$$

This concludes that $\rho_{\mathbf{X},\mathbf{Y}}$ and $\rho_{\mathbf{X},\mathbf{Z}}$ are also independent. ■

Recall that the derivation of Eq. A.15 uses the generalized central limit theorem which requires the limiting condition $q \to \infty$. Therefore it is important to check how the dimensionality q effects the quality of the prediction.

Fig. 2 displays the empirical distribution of ρ, that is $g(\rho) = \sum\limits_{1 \leq i < j \leq n} \delta_{\rho_{ij}}$, and the theoretical prediction in Eq. A.15 for various dimensions q. For the simulation, $n = 10000$ random vectors are drawn independently from $\mathcal{C}(0,1)$. We use the leading orders of the Laurent series of Eq. A.15 to represent the theoretical predictions.

It can be seen that for a sufficiently high-dimensional space, say $q = 2000$, the theoretical prediction fits the simulation very well. Moreover, the pairwise angles among Cauchy random vectors converge to $\frac{\pi}{2}$ as the dimensionality increases.

It implies that in high-dimensional spaces the distributional uniformity of normalized Cauchy random vectors could be tenable. We explain this in an intuitive way. According to Remark A 2, each element in the normalized variable converges independently in distribution to a Dirac δ-function, which can be constructed as the limit of a sequence of zero-centered normal distribution

$$f_{X_i'}(x_i') = \frac{1}{a\sqrt{\pi}}\mathrm{e}^{-\frac{x_i'^2}{a^2}} \quad \text{for} \quad a \to 0^+.$$

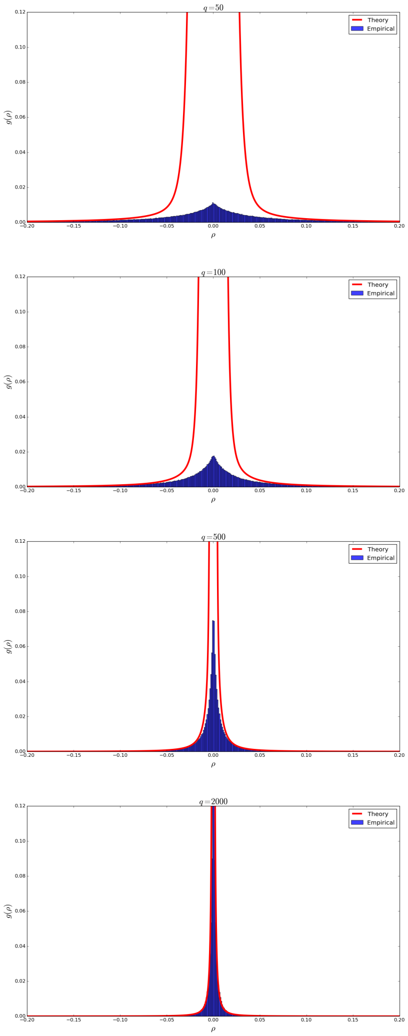


Figure 2: Comparisons between empirical distributions and theoretical predictions of ρ_C for various dimensions, $q = 50, 100, 500, 2000$.

Thus, following Lemma A 3, the density function of $f_{\mathbf{X}'}(x'_1, \cdots, x'_q)$ can be approximated by

$$f_{\mathbf{X}'}(x'_1, \cdots, x'_q) = \left(\frac{1}{a\sqrt{\pi}}\right)^q \mathrm{e}^{-\frac{\mathbf{x}'^\intercal \mathbf{x}'}{a^2}} \quad \text{for} \quad a \to 0^+.$$

This joint distribution is invariant under an arbitrary orthogonal rotation. Thus, it is a spherical distribution, as well as a uniform distribution on S^{q-1}. A rigorous proof of this result is still necessary. However, it is beyond the scope of this work.

A.3 DERIVATION OF COROLLARY 3

Corollary 3. *Consider a set of independent q-dimensional Cauchy random vectors which are pairwise ϵ-orthogonal with probability $1-\nu$. Then the number of such Cauchy random vectors is bounded by*

$$N \leq \sqrt{\frac{\pi\epsilon q}{4}} \left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}. \tag{A.16}$$

Proof. The derivation of this bound is similar to that of Corollary 2. The probability, that two random vectors whose elements are independently and identically Cauchy distributed are not ϵ-orthogonal, is bounded from above by

$$\Pr(|\rho| \geq \epsilon) = 2\int_\epsilon^1 \frac{2}{\pi q \rho^2}\,\mathrm{d}\rho < \frac{4}{\pi q}\frac{1}{\epsilon},$$

where only the leading order Laurent expansion of Eq. A.15 is considered. Then the quantity $\mathcal{P}(\epsilon, N)$ can be estimated as follows,

$$\mathcal{P}(\epsilon, N) := \prod_{k=1}^{N-1}\left[1 - k\Pr(|\rho| \geq \epsilon)\right] > \prod_{k=1}^{N-1}(1 - k\frac{4}{\pi\epsilon q})$$
$$> (1 - N\frac{4}{\pi\epsilon q})^N \sim \mathrm{e}^{-N^2\frac{4}{\pi\epsilon q}},$$

for sufficiently large N, and $q \to \infty$, with $N\frac{4}{\pi\epsilon q} < 1$. If we require $\mathcal{P}(\epsilon, N) \geq 1-\nu$, then the number of pairwise ϵ-orthogonal i.i.d. Cauchy random vectors is upper bounded by

$$\mathrm{e}^{-N^2\frac{4}{\pi\epsilon q}} \geq 1-\nu \;\Rightarrow\; N \leq \sqrt{\frac{\pi\epsilon q}{4}}\left[\log\left(\frac{1}{1-\nu}\right)\right]^{\frac{1}{2}}$$

■

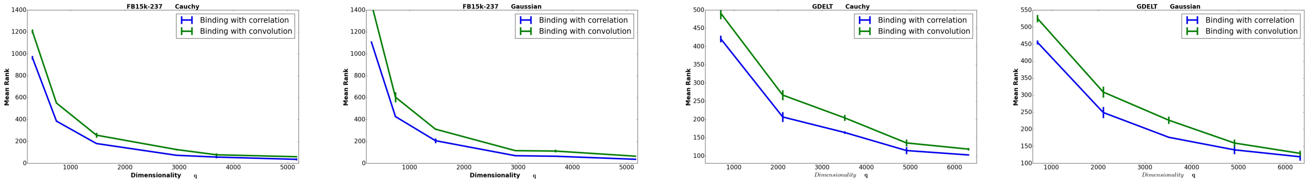
A.4 BINDING WITH CORRELATION OR CONVOLUTION

The filtered mean rank scores with different binding operations are compared in Fig. 3.

Now we give a heuristic explanation. For the sake of simplicity, consider only one semantic triple (s, p, o). For the binding with circular correlation the holistic representations are given by $\mathbf{h}_s^\mathrm{corr} = \mathbf{r}_p \star \mathbf{r}_o + \xi\mathbf{r}_s$, $\mathbf{h}_p^\mathrm{corr} = \mathbf{r}_s \star \mathbf{r}_o + \xi\mathbf{r}_p$, and $\mathbf{h}_o^\mathrm{corr} = \mathbf{r}_p \star \mathbf{r}_s + \xi\mathbf{r}_o$.

On the other hand, for the binding with convolution, the holistic representations given by: $\mathbf{h}_s^\mathrm{conv} = \mathbf{r}_p * \mathbf{r}_o + \xi\mathbf{r}_s$, $\mathbf{h}_p^\mathrm{conv} = \mathbf{r}_s * \mathbf{r}_o + \xi\mathbf{r}_p$, and $\mathbf{h}_o^\mathrm{conv} = \mathbf{r}_p * \mathbf{r}_s + \xi\mathbf{r}_o$.

Suppose that the subject needs to be retrieved and recalled using holistic representations only. To quantify the retrieval quality, a similarity $s^{\mathrm{corr}/\mathrm{conv}}$ is introduced for different binding operators. In particular, for binding with circular correlation $s^\mathrm{corr} := \mathbf{h}_s^{\mathrm{corr}\,\intercal}(\mathbf{h}_p^\mathrm{corr} * \mathbf{h}_o^\mathrm{corr})$,



(a) (b) (c) (d)

Figure 3: Comparison of the filtered MR scores for binding with convolution and binding with correlation (a) for FB15k-237 with Cauchy initialization, (b) for FB15k-237 with Gaussian initialization, (c) for GDELT dataset with Cauchy initialization, (d) for GDELT with Gaussian initialization

while for binding with circular convolution $s^{\text{conv}} := \mathbf{h}_s^{\text{conv}\,\intercal}(\mathbf{h}_p^{\text{conv}} \star \mathbf{h}_o^{\text{conv}})$.

Before any further derivations, recall that circular correlation can be computed in log-linear complexity via

$$\mathbf{a} \star \mathbf{b} = \mathcal{F}^{-1}\left(\overline{\mathcal{F}(\mathbf{a})} \odot \mathcal{F}(\mathbf{b})\right),$$

where $\mathcal{F}(\cdot)$ denotes the *fast Fourier transform* and $\mathcal{F}^{-1}(\cdot)$ its inverse, and the bar denotes the complex conjugate of a complex-valued vector. Moreover, circular convolution can also be computed via *fast Fourier transforms*

$$\mathbf{a} * \mathbf{b} = \mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{a}) \odot \mathcal{F}(\mathbf{b})\right).$$

First we compute the similarity s^{corr}

$$\begin{aligned}
s^{\text{corr}} &= \mathbf{h}_s^{\text{corr}\,\intercal}(\mathbf{h}_p^{\text{corr}} * \mathbf{h}_o^{\text{corr}}) \\
&= (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [(\mathbf{r}_s \star \mathbf{r}_o + \xi \mathbf{r}_p) * (\mathbf{r}_p \star \mathbf{r}_s + \xi \mathbf{r}_o)] \\
&= (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [\underbrace{(\mathbf{r}_s \star \mathbf{r}_o) * (\mathbf{r}_p \star \mathbf{r}_s)}_{\textcircled{1}} + \\
&\qquad \xi \underbrace{(\mathbf{r}_s \star \mathbf{r}_o) * \mathbf{r}_o}_{\textcircled{2}} + \xi \underbrace{\mathbf{r}_p * (\mathbf{r}_p \star \mathbf{r}_s)}_{\textcircled{3}} + \xi^2 \mathbf{r}_p * \mathbf{r}_o].
\end{aligned}$$

Using that

$$\begin{aligned}
\textcircled{1} &= \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \mathcal{F}(\mathbf{r}_o) \odot \overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_p \star \mathbf{r}_o, \\
\textcircled{2} &= \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \mathcal{F}(\mathbf{r}_o) \odot \mathcal{F}(\mathbf{r}_o)\right] = \text{Noise}, \\
\textcircled{3} &= \mathcal{F}^{-1}\left[\mathcal{F}(\mathbf{r}_p) \odot \overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_s,
\end{aligned}$$

yields

$$\begin{aligned}
s^{\text{corr}} &\approx (\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [\mathbf{r}_p \star \mathbf{r}_o + \xi \mathbf{r}_s + \text{Noise}] \\
&\approx (1 + \xi^2) + \text{Noise}.
\end{aligned}$$

The similarity s^{conv} can be computed in a similar way,

$$\begin{aligned}
s^{\text{conv}} &= \mathbf{h}_s^{\text{conv}\,\intercal}(\mathbf{h}_p^{\text{conv}} \star \mathbf{h}_o^{\text{conv}}) \\
&= (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [(\mathbf{r}_s * \mathbf{r}_o + \xi \mathbf{r}_p) \star (\mathbf{r}_p * \mathbf{r}_s + \xi \mathbf{r}_o)] \\
&= (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [\underbrace{(\mathbf{r}_s * \mathbf{r}_o) \star (\mathbf{r}_p * \mathbf{r}_s)}_{\textcircled{1}} + \\
&\qquad \xi \underbrace{(\mathbf{r}_s * \mathbf{r}_o) \star \mathbf{r}_o}_{\textcircled{2}} + \xi \underbrace{\mathbf{r}_p \star (\mathbf{r}_p * \mathbf{r}_s)}_{\textcircled{3}} + \xi^2 \mathbf{r}_p \star \mathbf{r}_o].
\end{aligned}$$

Moreover, using that

$$\begin{aligned}
\textcircled{1} &= \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \overline{\mathcal{F}(\mathbf{r}_o)} \odot \mathcal{F}(\mathbf{r}_p) \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_o \star \mathbf{r}_p, \\
\textcircled{2} &= \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_s)} \odot \overline{\mathcal{F}(\mathbf{r}_o)} \odot \mathcal{F}(\mathbf{r}_o)\right] \approx \mathbf{r}_s, \\
\textcircled{3} &= \mathcal{F}^{-1}\left[\overline{\mathcal{F}(\mathbf{r}_p)} \odot \mathcal{F}(\mathbf{r}_p) \odot \mathcal{F}(\mathbf{r}_s)\right] \approx \mathbf{r}_s,
\end{aligned}$$

leads to

$$\begin{aligned}
s^{\text{conv}} &\approx (\mathbf{r}_p * \mathbf{r}_o + \xi \mathbf{r}_s)^\intercal [\mathbf{r}_o \star \mathbf{r}_p + 2\xi \mathbf{r}_s + \text{Noise}] \\
&\approx 2\xi^2 + \text{Noise}.
\end{aligned}$$

The optimal hyper-parameter requires $\xi < 1$ which in turn yields $s^{\text{corr}} > s^{\text{conv}}$. From the derivation of s^{corr}, we have that the subject-object association pair stored in $\mathbf{h}_p^{\text{corr}}$ contributes the most in $s^{\text{corr}} \approx 1 + \xi^2$ via the term $\textcircled{1}$.

A.5 APPROXIMATION OF $\rho_{\mathbf{r}'_{\mathrm{o}},\mathbf{h}_{\mathrm{o}}}$

Here we provide a heuristic study on the relations between hyper-parameter ξ, $\lambda_{\mathrm{G/C}}$, and the average number of association pairs N_{a}. Recall that ξ was introduced for holistic representations, and $\lambda_{\mathrm{G/C}}$ is defined as $\lambda_{\mathrm{G/C}} := \mathbb{E}[|\rho_{\mathrm{G/C}}|]$.

Consider a subject s. The predicate-object pair (p, o) is stored in the holistic representation $\mathbf{h}_{\mathrm{s}}$ along with the other $N_{\mathrm{a}} - 1$ pairs. This means

$$\mathbf{h}_{\mathrm{s}} = \xi N_{\mathrm{a}} \mathbf{r}_{\mathrm{s}} + \mathbf{r}_{\mathrm{p}} \star \mathbf{r}_{\mathrm{o}} + \sum_{i=2}^{N_{\mathrm{a}}} \mathbf{r}_{p_i} \star \mathbf{r}_{o_i}.$$

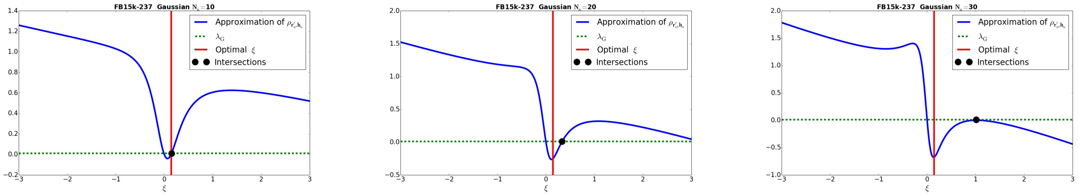


Figure 4: Approximations of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}(\xi, N_\mathrm{a})$ in the case of Gaussian holistic representations with (a): $N_\mathrm{a} = 10$ (b): $N_\mathrm{a} = 20$ (c): $N_\mathrm{a} = 30$. We use the experiment setting with dimnsionality $q = 5200$, $\lambda_\mathrm{G} = 0.0111$, and optimal $\xi = 0.14$.

Suppose that we aim to identify the object in the triple $(\mathrm{s}, \mathrm{p}, \cdot)$ via $\mathbf{h}_\mathrm{s}$ and $\mathbf{h}_\mathrm{p}$, where $\mathbf{h}_\mathrm{p}$ is the holistic representation for the predicate p. We further assume that up to N_a subject-object pairs can be stored in $\mathbf{h}_\mathrm{p}$ having high enough fidelity, then

$$\mathbf{h}_\mathrm{p} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{p} + \sum_{k=1}^{N_\mathrm{a}} \mathbf{r}_{s_k} \star \mathbf{r}_{o_k}.$$

To retrieve the object o, the decoding via circular convolution is obtained as follows

$$\begin{aligned}
&\mathbf{r}'_\mathrm{o} = \mathbf{h}_\mathrm{p} * \mathbf{h}_\mathrm{s}\\
&\approx \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \xi^2 N_\mathrm{a}^2(\mathbf{r}_\mathrm{p} * \mathbf{r}_\mathrm{s}) + \xi N_\mathrm{a}\sum_{i=2}^{N_\mathrm{a}}[\mathbf{r}_\mathrm{p} * (\mathbf{r}_{p_i} \star \mathbf{r}_{o_i})]\\
&+ \xi N_\mathrm{a}\sum_{k=1}^{N_\mathrm{a}}[(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * \mathbf{r}_\mathrm{s}] + \sum_{k=1}^{N_\mathrm{a}}[(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * (\mathbf{r}_\mathrm{p} \star \mathbf{r}_\mathrm{o})]\\
&+ \sum_{k=1,i=2}^{N_\mathrm{a},N_\mathrm{a}}[(\mathbf{r}_{s_k} \star \mathbf{r}_{o_k}) * (\mathbf{r}_{p_i} \star \mathbf{r}_{o_i})]\\
&= \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \xi^2 N_\mathrm{a}^2\mathbf{b}_1 + \xi N_\mathrm{a}\sum_{i=2}^{N_\mathrm{a}}\mathbf{b}_i + \xi N_\mathrm{a}\sum_{k=1}^{N_\mathrm{a}}\mathbf{c}_k\\
&+ \sum_{k=1}^{N_\mathrm{a}}\mathbf{d}_k + \sum_{k=1,i=2}^{N_\mathrm{a},N_\mathrm{a}}\mathbf{e}_{ki},
\end{aligned}$$

where $\mathbf{b}_i$, $\mathbf{c}_k$, $\mathbf{d}_k$, and $\mathbf{e}_{ki}$ with $i, k = 1, \cdots, N_\mathrm{a}$ are approximately normalized Gaussian/Cauchy random vectors. This is due to the fact that in high-dimensional spaces both circular correlation and circular convolution of two normalized Gaussian/Cauchy random vectors is approximately a normalized Gaussian/Cauchy random vectors.

After decoding with circular convolutions, the decoded noisy version of the object needs to be recalled with $\mathbf{h}_\mathrm{o}$ which is the holistic representation of o. As before, N_a predicate-subject association pairs are assumed to be stored in the holistic representation of o, with

$$\mathbf{h}_\mathrm{o} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \sum_{j=1}^{N_\mathrm{a}} \mathbf{r}_{p_j} \star \mathbf{r}_{s_j} = \xi N_\mathrm{a}\mathbf{r}_\mathrm{o} + \sum_{j=1}^{N_\mathrm{a}} \mathbf{f}_j,$$

where $\mathbf{f}_j$, $j = 1, \cdots, N_\mathrm{a}$ are approximately normalized Gaussian/Cauchy random vectors.

In order to recall the object successfully, the angle between $\mathbf{r}'_\mathrm{o}$ and $\mathbf{h_o}$ should be smaller than the expected absolute angle between two arbitrary vectors, namely $\theta_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}} < \mathbb{E}[|\theta_{\mathrm{G/C}}|]$. Given the definition of λ, equivalently, it requires $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}} > \lambda_{\mathrm{G/C}}$.

Now we turn to approximate the numerator of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}$, that is $\mathbf{r}'^\intercal_\mathrm{o}\mathbf{h}_\mathrm{o}$. Recall that, in general, the expectation of the dot product of two normalized, independent random vectors equals 0 due to the symmetry of the density function $g(\rho_{\mathrm{G/C}})$. Therefore, in the following approximation we only consider noisy terms which are directly related to $\mathbf{r}_\mathrm{o}$ as adverse effects to a successful retrieval and treat other terms as white noisy with zero expectation. This yields,

$$\begin{aligned}
&\mathbf{r}'^\intercal_\mathrm{o}\mathbf{h}_\mathrm{o}\\
&\approx \xi^2 N_\mathrm{a}^2 + \xi N_\mathrm{a}\sum_{j=1}^{N_\mathrm{a}}(\mathbf{r}_\mathrm{o}^\intercal\mathbf{f}_j) + \xi^3 N_\mathrm{a}^3(\mathbf{r}_\mathrm{o}^\intercal\mathbf{b}_1) + \xi^2 N_\mathrm{a}^2\sum_{i=2}^{N_\mathrm{a}}(\mathbf{r}_\mathrm{o}^\intercal\mathbf{b}_i)\\
&\quad + \xi^2 N_\mathrm{a}^2\sum_{k=1}^{N_\mathrm{a}}(\mathbf{r}_\mathrm{o}^\intercal\mathbf{c}_k) + \xi N_\mathrm{a}\sum_{k=1}^{N_\mathrm{a}}(\mathbf{r}_\mathrm{o}^\intercal\mathbf{d}_k) + \xi N_\mathrm{a}\sum_{k=1,i=2}^{N_\mathrm{a},N_\mathrm{a}}(\mathbf{r}_\mathrm{o}^\intercal\mathbf{e}_{ki})\\
&> \xi^2 N_\mathrm{a}^2 - (\xi N_\mathrm{a}^2 + \xi^3 N_\mathrm{a}^3 + \xi^2 N_\mathrm{a}^2(N_\mathrm{a} - 1) + \xi^2 N_\mathrm{a}^3\\
&\quad + \xi N_\mathrm{a}^2 + \xi N_\mathrm{a}^2(N_\mathrm{a} - 1))\lambda_{\mathrm{G/C}}\\
&= \xi^2 N_\mathrm{a}^2 - (\xi^3 N_\mathrm{a}^3 + 2\xi^2 N_\mathrm{a}^3 - \xi^2 N_\mathrm{a}^2 + \xi N_\mathrm{a}^2 + \xi N_\mathrm{a}^3)\lambda_{\mathrm{G/C}}.
\end{aligned}$$

Furthermore, the denominator of $\rho_{\mathbf{r}'_\mathrm{o},\mathbf{h}_\mathrm{o}}$ can be approximated in the same way. More concretely, we have

$$\begin{aligned}
||\mathbf{r}'_\mathrm{o}|| \cdot ||\mathbf{h}_\mathrm{o}|| < \xi^2 N_\mathrm{a}^2 + N_\mathrm{a} + 2\xi N_\mathrm{a}^2\lambda_{\mathrm{G/C}}\\
+ N_\mathrm{a}(N_\mathrm{a} - 1)\lambda_{\mathrm{G/C}}.
\end{aligned}$$

Combining these results, a sufficient condition to retrieve the object correctly is given by

$$\rho_{\mathbf{r}'_o,\mathbf{h}_o} > \frac{\xi^2 N_a^2 - (\xi^3 N_a^3 + 2\xi^2 N_a^3 - \xi^2 N_a^2 + \xi N_a^2 + \xi N_a^3)\lambda_{G/C}}{\xi^2 N_a^2 + N_a + 2\xi N_a^2 \lambda_{G/C} + N_a(N_a - 1)\lambda_{G/C}} > \lambda_{G/C}. \tag{A.17}$$

Consider the experimental setting for the memorization task on the FB15k-237 dataset: The dimensionality of the holistic representations is $q = 5200$, $\lambda_G(q = 5200) = 0.0111$, and $\lambda_C(q = 5200) = 0.00204$. Fig. 4 displays the above approximation of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a)$ for Gaussian initializations.

After performing grid search, the optimal ξ is found to be close to the intersection of the curve $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a = 10)$ and the threshold λ_G. However, for $N_a > 30$, no intersection points on $\xi > 0$ exists. This explains why Gaussian holistic representations have lower memory capacity compared to Cauchy holistic representations.

More comparisons between Gaussian and Cauchy initializations can be found in Fig. 5.

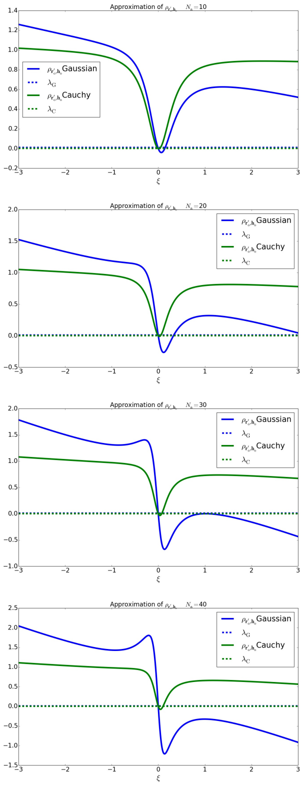


Figure 5: Comparison of $\rho_{\mathbf{r}'_o,\mathbf{h}_o}(\xi, N_a)$ for Gaussian (blue) and Cauchy (green) holistic representations with (a): $N_a = 10$ (b): $N_a = 20$ (c): $N_a = 30$ (d): $N_a = 40$.

A.6 HOLISTIC ENCODING ALGORITHM

Algorithm 1 Holistic Encoding

```
Require: hyper-parameter $\xi$
for $i = 1, \cdots, N_e$ do
    Draw $\tilde{\mathbf{r}}_{e_i}^{G/C}$ from Gaussian or Cauchy
    $\mathbf{r}_{e_i}^{G/C} \leftarrow \mathbf{Norm}(\tilde{\mathbf{r}}_{e_i}^{G/C})$
for $i = 1, \cdots, N_p$ do
    Draw $\tilde{\mathbf{r}}_{p_i}^{G/C}$ from Gaussian or Cauchy
    $\mathbf{r}_{p_i}^{G/C} \leftarrow \mathbf{Norm}(\tilde{\mathbf{r}}_{p_i}^{G/C})$
for $i = 1, \cdots, N_e$ do
    Extract $\in \mathcal{S}^s(e_i), \mathcal{S}^o(e_i)$ from Database
    $\mathbf{h}_{e_i}^s \leftarrow \sum_{(p,o)\in\mathcal{S}^s(e_i)} [\mathbf{Norm}(\mathbf{r}_p \star \mathbf{r}_o) + \xi \mathbf{r}_{e_i}]$
    $\mathbf{h}_{e_i}^o \leftarrow \sum_{(s,p)\in\mathcal{S}^o(e_i)} [\mathbf{Norm}(\mathbf{r}_p \star \mathbf{r}_s) + \xi \mathbf{r}_{e_i}]$
    $\mathbf{h}_{e_i} \leftarrow \mathbf{h}_{e_i}^s + \mathbf{h}_{e_i}^o$
for $i = 1, \cdots, N_p$ do
    Extract $\mathcal{S}(p_i)$ from Database
    $\mathbf{h}_{p_i} \leftarrow \sum_{(s,o)\in\mathcal{S}(p_i)} [\mathbf{Norm}(\mathbf{r}_s \star \mathbf{r}_o) + \xi \mathbf{r}_{p_i}]$
```

Remark:

Normalizing initial random vectors can assist the analysis of memory capacities via different sampling schemes. For example, for the derivation of retrieval condition Eq. A.17 we heavily relay on the fact that the dot product of two random vectors - say $\mathbf{r}_i \cdot \mathbf{r}_j$, where $\mathbf{r}_i$ and $\mathbf{r}_j$ are randomly sampled and normalized - is just ρ_{ij}. In the memorization task, since triples are recalled by comparing the angles (a.k.a cosine similarity) between decoded noisy vector and all other holistic vectors, normalization does not effect the recall scores.

A.7 NOTATIONS

In Table 1 and Table 2, we summary important notations introduced in Section 3 and 4, respectively.

A.8 FURTHER EXPERIMENTAL DETAILS

After searching for the optimal hyper-parameter ξ for holistic encoding, holistic representations with superior memory capacity will be fixed and applied to the next inference tasks.

The architecture is a simple 2-layered fully-connected neural network, which map high-dimensional holistic representations ($q = 3600$) of subjects, predicates, and objects to low-dimensional ($h_2 = 256$) representations, separately. We choose ReLU as the activation function for faster training, and batch normalization after the hidden-layer for regularization. In order to reduce the

Table 1: Notations for ϵ-orthogonality

Symbol	Meaning
$\mathbf{X}$	q-dimensional random variable with elements drawn from Gaussian or Cauchy distribution
Θ_{ij}	Angle between two random variables $\mathbf{X}_i$ and $\mathbf{X}_j$
ρ_{ij}	Cosine of the angle between random variables $\mathbf{X}_i$ and $\mathbf{X}_j$
$g(\rho_{\mathrm{G}})$	Asymptotic density function of ρ_{ij} given an ensemble of Gaussian random variables $\mathbf{X}_i$, $i = 1, \cdots, n$, with $n \to \infty$
$g(\rho_{\mathrm{C}})$	Asymptotic density function of ρ_{ij} given an ensemble of Cauchy random variables $\mathbf{X}_i$, $i = 1, \cdots, n$, with $n \to \infty$
λ_{G}	Expectation value of $\|\rho_{\mathrm{G}}\|$
λ_{C}	Expectation value of $\|\rho_{\mathrm{C}}\|$

number of trainable parameters, the network has a bottleneck structure with the dimensionality of the hidden-layer $h_1 = 64$. The extracted low-dimensional features are then combined via tri-linear dot-product, similar to DISTMULT.

In summary, given a triple $(\mathrm{s}, \mathrm{p}, \mathrm{o})$ the scoring function η_{spo} takes the following form:

$$\begin{aligned}\eta_{\mathrm{spo}} = &\langle \mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{s}}\mathbf{W}_1^e))\mathbf{W}_2^e, \\ &\mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{p}}\mathbf{W}_1^p))\mathbf{W}_2^p, \\ &\mathrm{BN}(\mathrm{ReLU}(\mathbf{h}_{\mathrm{o}}\mathbf{W}_1^e))\mathbf{W}_2^e\rangle,\end{aligned}$$

where $\mathbf{h}_{\mathrm{s}}$, $\mathbf{h}_{\mathrm{s}}$ are the holistic representations for the subject s and object o; $\mathbf{h}_{\mathrm{p}}$ is the holistic representation for the predicate p. Note that there are two separate networks for extracting low-dimensional features of entities and predicates, respectively. In particular, $\mathbf{W}_1^e \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^e \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for entities, including subjects and objects; $\mathbf{W}_1^p \in \mathbb{R}^{q \times h_1}$ and $\mathbf{W}_2^p \in \mathbb{R}^{h_1 \times h_2}$ are shared weights for predicates.

For training the model, we minimize the following binary cross-entropy loss with l_2 regularization:

$$\begin{aligned}\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}&(y_i \cdot \log(\sigma(\eta_{x_i})) + \\ &(1 - y_i) \cdot \log(1 - \sigma(\eta_{x_i}))) + \lambda||\mathcal{A}||_2^2,\end{aligned}$$

where the label vector y_i has dimension $\{0, 1\}^{1 \times N}$ for 1-N scoring to accelerate the link prediction tasks. To be more specific, during the training given a triple $(\mathrm{s}, \mathrm{p}, \mathrm{o})$, we take the subject-predicate pair (s, p) and and rank it against all object entities $o \in \mathcal{E}$; take the predicate-object pair (p, o) and rank it against all subject entities $s \in \mathcal{E}$ simultaneously as well.

Table 2: Notations for holistic representations

Symbol	Meaning
$*$	Circular convolution
$\star$	Circular correlation
Norm	Normalization operator, $\mathrm{Norm}(\mathbf{r}) := \frac{\mathbf{r}}{\|\|\mathbf{r}\|\|}$
N_e	Number of entities in the KG
N_p	Number of predicates in the KG
N_a	Average number of association pairs encoded in holistic representations of entities
$\mathbf{r}_{e_i}^{\mathrm{G/C}}$	Random initialization of entity e_i with elements drawn from Gaussian or Cauchy distribution
$\mathbf{r}_{p_i}^{\mathrm{G/C}}$	Random initialization of predicate p_i with elements drawn from Gaussian or Cauchy distribution
$\mathbf{h}_{e_i}^s$	Holistic representation of entity e_i as subject
$\mathbf{h}_{e_i}^o$	Holistic representation of entity e_i as object
$\mathbf{h}_{e_i}$	Overall holistic representation of entity e_i
$\mathbf{h}_{p_i}$	Holistic representation of predicate p_i
ξ	Hyper-parameter for holistic encoding

Hyper-parameters in the HOLNN$_{\mathrm{G}}$ and HOLNN$_{\mathrm{C}}$ are optimized via grid search with respect to the mean reciprocal rank (MRR). The ranges for grid search are as follows - learning rate $\{0.001, 0.003, 0.005\}$, $l2$ regularization parameter $\{0., 0.01, 0.05\}$, decay parameter in the batch normalization $\{0.99, 0.9, 0.8, 0.7\}$, and batch size $\{1000, 3000, 5000\}$.

References

Cai, T Tony and Tiefeng Jiang (2012). "Phase transition in limiting distributions of coherence of high-dimensional random matrices". In: *Journal of Multivariate Analysis* 107, pp. 24–39.

Gnedenko, B.V. and A.N. Kolmogorov (1954). *Limit distributions for sums of independent random variables*. Addison-Wesley. URL: https://books.google.de/books?id=7qVyAQAACAAJ.

Mandelbrot, Benoit (1960). "The Pareto-Levy law and the distribution of income". In: *International Economic Review* 1.2, pp. 79–106.

Muirhead, Robb J (2009). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Nolan, John (2003). *Stable distributions: models for heavy-tailed data*. Birkhauser New York.