**Knowledge Graph Embedding: A Probabilistic Perspective and Generalization Bounds**

Anonymous authors  
Paper under double-blind review

**Abstract**

We study theoretical properties of embedding methods for knowledge graph completion under the “missing completely at random” assumption. We prove generalization error bounds for this setting. Even though the missing completely at random setting may seem naive, it is actually how knowledge graph embedding methods are typically benchmarked in the literature. Our results provide, to certain extent, an explanation for why knowledge graph embedding methods work (as much as classical learning theory results provide explanations for classical learning from i.i.d. data).

1 Introduction

Knowledge bases (also called knowledge graphs) such as YAGO (Mahdisoltani et al., 2015) and NELL (Carlson et al., 2010) contain facts in the form of triples (object, relation, object), e.g. (NewYork, IsLocatedIn, US). Most knowledge bases are incomplete, i.e. there are facts that should be included in them but are not. A popular task is to automatically complete such knowledge bases. This is often done by embedding the objects and relations contained in them into some vector spaces and then using the embedding to predict the missing tuples. There has been a number of such methods presented in the literature (we give an overview of existing methods in Section 2.3). However, to our best knowledge, there has not been a theoretical analysis of the generalization abilities of these methods in the spirit of classical learning theory. We make the first steps in this direction in the present paper. We derive expected error bounds for a class of knowledge graph embedding methods.

A key challenge in deriving generalization bounds in this setting is that the training and test data do not satisfy assumptions typically used in learning theory. The knowledge bases that we receive contain only positive examples, that is, the facts already in the knowledge base. We do not know which of the facts not contained in the knowledge base are just missing and which are not contained in the knowledge base because they are actually false. This is known as learning from positive and unlabeled data (PU-learning) so far; we refer to (Bekker & Davis, 2018) for a recent survey. Only a few results have been obtained for generalization bounds for PU-learning, e.g. (Niu et al., 2016), but they are not directly applicable to our setting (we discuss the details in Section 10).

The bounds that we derive in this paper rely on an assumption on how the triples are missing from the knowledge graph. In particular, we assume that the triples are missing “completely at random”, i.e. that each tuple may be missing with probability $\delta > 0$ independently of the others. Before dismissing this setting as too naive, please bare in mind that this is actually exactly how some famous datasets’ training and test sets are created, e.g. (Bordes et al., 2013). Thus, the theoretical results actually do relate to what one observes in “practice” (i.e. in the respective papers that use these datasets).

2 Preliminaries

2.1 Knowledge Graphs

Let $O$ be a set of objects and $R$ a set of relations. A knowledge graph is a set of triples $(h, r, t)$ where $h$ and $t$ are objects and $r$ is a relation. Here $h$ is called the head of the triple and $t$ is called the tail of the triple. In the formalism of first-order logic, a triple $(h, r, t)$ represents a first-order logic
atom \( r(h, t) \). For instance the triple (Alice, Knows, Bob) corresponds to the first-order logic atom Knows(Alice, Bob), which represents the fact that Alice knows Bob. Here, Alice and Bob are objects and Knows is a relation.

**Example 1.** Let us have the following objects and relations: \( O = \{ \text{France, Germany, Berlin, Paris} \} \), \( R = \{ \text{CapitalOf, Borders} \} \). Then the knowledge graph, representing our geographical knowledge about these countries and cities, can look as \( G = \{ \{ \text{France, Borders, Germany} \}, \{ \text{Germany, Borders, France} \}, \{ \text{Berlin, CapitalOf, Germany} \}, \{ \text{Paris, CapitalOf, France} \} \} \). Whereas, in this case, the knowledge graph is complete and accurate, in reality most knowledge graphs will lack some of the true tuples. The task of knowledge base completion is to add such missing facts.

### 2.2 Knowledge Graph Embedding

Knowledge graph embedding methods represent knowledge graphs geometrically. They do so by representing objects and relations as vectors (or matrices). Given a knowledge graph \( G \) on some sets of objects \( O \) and relations \( R \), most knowledge graph embedding methods need three components: vector representations of objects, vector representations of relations and a function \( \psi(x_h, x_r, x_t) \) which takes the representations of \( h, r \) and \( t \) and produces a score that is used to assess how likely it is that the relation \((h, r, t)\) is contained in the knowledge graph. One can then compare the output of the function with a threshold to decide which triples are (predicted to be) in the knowledge graph.

**Note on notation.** For a set of objects \( O \) and a set of relations \( R \) we will normally denote their vector representations by \( X \). We will usually treat \( X \) as a look-up table. For an object \( o \) we will typically use \( x_o \) to denote its vector representation from \( X \) and analogically for a relation \( r \) we will use \( x_r \), unless explicitly stated otherwise. With a slight abuse of notation, we will also sometimes treat \( X \) as a set of vectors and, for instance, write \( x \in X \). Relations can be mapped to matrices (denoted \( M_r \)) or tensors instead of only vectors, and even if relations are mapped to vectors the dimension may be different from the dimension of entity embeddings. However, for the sake of simplicity we will not introduce more notations, and \( X \) and \( \psi \) will also be used in these cases.

### 2.3 Existing Methods

One can roughly categorize KG embedding approaches into several groups based on different scoring functions such as distance-based scoring functions and similarity-based scoring functions.

The most well-known class of embedding methods using distance-based scoring functions is TransE (Bordes et al. [2013]) and its variants. The scoring function of TransE is defined as the negative distance between \( x_h + x_r \) and \( x_t \), i.e., 
\[
\psi(x_h, x_r, x_t) = -||x_h + x_r - x_t||_2.
\]

In order to effectively model more general relations rather than only one-to-one relations, TransH has been proposed (Wang et al. [2014]). Given a triple \((h, r, t)\), the scoring function is defined as 
\[
-||x_h^+ + x_r - x_t^+||_2
\]
where \( x_h^+ \) and \( x_t^+ \) are the projections of entity embeddings \( x_h \) and \( x_t \) onto some hyperplane, respectively. TransR (Lin et al., 2015), TransD (Ji et al., 2015) and TranSparse (Ji et al., 2016) are similar to TransH, allowing relation-specific embeddings in different ways. TransM (Fan et al., 2014) and ManifoldE (Xiao et al., 2015) relax the requirement \( x_h + x_r \approx x_t \).

SE (Bordes et al., 2011) uses two matrices to project head and tail entities for each relation \( r \), and the score is 
\[
-||M_r^2x_h - M_r^2x_t||_1.
\]

RotatE (Sun et al., 2019) defines each relation as a rotation from the source entity to the target entity in the complex vector space, and its scoring function is 
\[
-||x_h \odot x_r - x_t||_2
\]
where \( \odot \) is the entry-wise product.

In the class of similarity-based methods, RESCAL (Nickel et al., 2011) and its variants are among the most well-known ones. RESCAL also assigns every entity a vector but every relation \( r \) a matrix \( M_r \). The scoring function is defined as \( x_h^\top M_r x_t \). Jenatton et al. (2012) further assumes that every \( M_r \) has the same set of eigenvectors. Garcia-Duran et al. (2014) adds other terms into the scoring function, which is typically used to denote its vector representation from \( X \).
function $x_h^\top M_r x_t + x_r^\top x_r + x_t^\top x_t + x_h^\top D x_h$ where $D$ is a diagonal matrix independent of relations. DistMult (Yang et al., 2015) is a simplified version of RESCAL by restricting $M_r$ being diagonal, and hence improves the efficiency of RESCAL. In order to combine the advantages of both methods, i.e., the expressive power of RESCAL and the simplicity of DistMult, HolE (Nickel et al., 2016) has been proposed. ComplEx (Yang et al., 2015) extends DistMult by allowing complex-valued embeddings, which is more powerful in modeling asymmetric relations. Hayashi & Shimbo (2017) show that ComplEx is in fact equivalent to HolE. In order to capture analogical structures, ANALOGY (Liu et al., 2017) further requires $M_r$ in RESCAL to be normal and mutually commutative.

A class of KG embeddings is based on tensor factorization approaches such as the Canonical Polyadic (CP) decomposition (Hitchcock, 1927) whose scoring function is similar to DistMult, but every entity has two independent embeddings. Simple (Kazemi & Poole, 2018) is a recent extension of CP decomposition, which ensures the two embeddings of every entity dependent.

There are also a group of algorithms using neural networks for embedding, e.g., SME (Bordes et al., 2014), NTN and SLM (Bordes et al., 2014), MLP (Dong et al., 2014), NAM (Liu et al., 2016) and ConvE (Dettmers et al., 2018).

All the aforementioned embedding methods map entities and relations in a deterministic way. In KG2E (He et al., 2015), representations of entities and relations are random vectors sampled from multivariate normal distributions,

$$x_h \sim \mathcal{N}(\mu_h, \Sigma_h), \quad x_r \sim \mathcal{N}(\mu_r, \Sigma_r), \quad x_t \sim \mathcal{N}(\mu_t, \Sigma_t).$$

A possible scoring function is $KL(x_t - x_h \| x_r)$ and another is the probability inner product (Jebara et al., 2004) of $x_t - x_h$ and $x_r$. TransG (Han et al., 2016) further models relations using mixtures of Gaussians.

## 3 Learning Setting

We work within the following learning setting. We assume that there exists some “ground truth” knowledge graph $G^*$ and that we observe only a subset of the triples contained in it. We will denote the observed knowledge graph consisting of a subset of $G^*$’s triples by $\hat{G}$. Furthermore, we assume that $\hat{G}$ is obtained from $G^*$ by the following sampling process, called data-generating distribution.

**Definition 1** (Data-generating distribution). Let $G^*$ be a ground-truth knowledge graph and $\delta > 0$ be a real number. The data-generating distribution is given by the following process. Iterate over all triples $\tau \in G^*$ and add each of them to $\hat{G}$ with probability $1 - \delta$ independently of the others. We call $\delta$ the missingness parameter.

The task that we want to solve is then to reconstruct the ground-truth knowledge graph $G^*$ as well as possible from a single sample from the data-generating distribution.

**Remark 1.** Some knowledge graph embedding methods use different representations for the same object depending on whether it is used as a head or tail of a particular triple. It is easy to see that our general setting also allows such representations; it is enough to double the dimension of the vectors and assume that the function $\psi(x_h, x_r, x_t)$ depends only on the first half of dimensions of $x_h$ and the second half of the dimensions of $x_t$. It is easy to see that we can also handle the case when the vector representations of relations have different dimension than those of objects similarly.

## 4 Intuition and Sketch of the Argument

Our main results are bounds for the expected number of incorrectly predicted triples for knowledge graph completion. In particular, we show that, assuming certain technical assumption on the knowledge graph embedding method and the way it is trained, there exists a constant $C$ such that the expected error rate in the realizable setting is bounded by $C \sqrt{|G^*|/|O|^2 \cdot \sqrt{\ln |O|}/|O|}$, where $G^*$ is the ground truth knowledge graph and $O$ is the set of objects. The first term in the product measures sparsity of the graph and the second term has the familiar form of classical error bounds for learning from i.i.d. data. We also provide similar bounds for the agnostic setting.
The strategy taken in this paper is based on viewing the problem as learning a distribution on knowledge graphs. Suppose that we have not only one sample \( \mathcal{G} \) (one knowledge graph) from the data-generating distribution, but actually many such samples \( \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \) (a collection of knowledge graphs instead of just one). Then we could produce a prediction for the ground-truth knowledge graph by taking the union \( \mathcal{G}_1 \cup \mathcal{G}_2 \cup \cdots \cup \mathcal{G}_n \). Another possibility would be to use the knowledge graphs \( \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n \) to estimate a distribution over knowledge graphs. Here, we could compute the empirical frequencies of the observed triples and predict those with frequency higher than some threshold \( t \) as true. However, in reality, we do not have a collection of knowledge graphs but only one. Therefore the first approach described above is not feasible. On the other hand, there is a hope for the second approach; we may still try to estimate the data-generating distribution only from one sample. To estimate the data-generating distribution, we can exploit the fact that we know its form and the independencies it satisfies. We define a distribution to approximate the data-generating distribution, parameterized by the vector representations \( \mathcal{X} \) of objects and relations.

For the approach sketched above to work, we need (i) the distribution to be expressive enough and (ii) we need to show that the learned distribution will be close enough to the data-generating distribution if we estimate it from a sufficiently large knowledge graph. The question of expressiveness has been tackled in the literature (e.g. Kazemi & Poole (2018)) and we do not deal with it much in this paper.

First we define log-likelihood and normalized log-likelihood of the KG-distribution model.

**Definition 2** (KG-distribution). Let \( \mathcal{O} \) be a set of objects, \( \mathcal{R} \) be a set of relations and \( \Omega = 2^{\mathcal{O} \times \mathcal{R} \times \mathcal{O}} \) be the set of all possible knowledge graphs on these objects and relations. Let \( \mathcal{G} \in \Omega \) be a knowledge graph and \( \mathcal{X} \) be the vector representation of the objects and relations. Then we define \( P_{\mathcal{X}}(\mathcal{G}) = \frac{1}{Z} \prod_{(h, r, t) \in \mathcal{G}} \exp (\psi(x_h, x_r, x_t)), \) where \( x_h, x_r, x_t \) are the vector representations of \( h, r \) and \( t \). Here, \( \psi: \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) is a function, \( d \) is the dimension of the vector representations, and finally \( Z = \sum_{\mathcal{G}' \in \Omega} \prod_{(h, r, t) \in \mathcal{G}'} \exp (\psi(x_h, x_r, x_t)) \) is a normalization constant, ensuring that \( P_{\mathcal{X}}(\mathcal{G}) \) is a probability distribution.

Unless we allow \( |\psi| \) to be infinite, KG-distributions cannot model the data-generating distribution exactly. However, as we will show it is enough if the KG-distributions are close enough in Kullback-Leibler divergence.

**6 Generalization Bounds for Log-Likelihood**

In this section we derive generalization bounds for log-likelihood of KG-distributions learned from one sample (i.e. knowledge graph) from the data-generating distribution described in Section 3.

These bounds are also valid when we sample from the noised data-generating distribution.

First we define log-likelihood and normalized log-likelihood of the KG-distribution model.

**Definition 3** (Normalized log-likelihood). Let \( \mathcal{G} \) be a knowledge graph on a set of objects \( \mathcal{O} \) and a set of relations \( \mathcal{R} \) and \( \mathcal{X} \) be vector representations of the objects and relations. The log-likelihood of this model is \( L(\mathcal{X} | \mathcal{G}) = \ln P_{\mathcal{X}}(\mathcal{G}) \) and its normalized log-likelihood is then defined as \( NL(\mathcal{X} | \mathcal{G}) = \frac{1}{|\mathcal{O}|^2 |\mathcal{R}|} \cdot L(\mathcal{X} | \mathcal{G}) \).

---

2 This naive approach would obviously fail if the data-generating process was also producing false positive errors, i.e. adding incorrect triples to the ground truth.
We start by proving a concentration inequality for the log-likelihood of KG-distributions, which is a rather straightforward consequence of McDiarmid’s inequality.

**Theorem 1.** Let $G^*$ be the ground-truth knowledge graph. Let $G$ be sampled from the data-generating distribution induced by $G^*$ and $X$ be fixed vector representations. Then

$$P \left( |L(X|G) - \mathbb{E}[L(X)]| \geq \epsilon \right) \leq 2 \exp \left( - \frac{2\epsilon^2}{M^2|G^*|} \right)$$

where $M \geq \sup_{h,t \in \mathcal{X}, r \in \mathcal{R}} |\psi(x_h, x_r, x_t)|$.

The next theorem is the main result of this section. It bounds the expectation of the maximum deviation of the log-likelihood of KG distributions from their mean, where the KG-distributions are selected from a hypothesis class $\mathcal{H}$ (which has to satisfy certain assumptions listed in the theorem).

**Theorem 2.** Let $G^*$ be a ground-truth knowledge graph on a set of objects $O$ and a set of relations $R$ and let $G$ be sampled by the data-generating distribution. Let $\mathcal{X} \subseteq \mathbb{R}^d$. Finally, let $\mathcal{H}_X$ (“hypothesis class”) denote the set of vector representations $X$, $X \subseteq \mathcal{X}^{(|O|+|R|)}$, of the objects from $O$ and relations from $R$. Let $D$, $M$ and $K$ be real numbers satisfying:

1. $\forall x, x' \in \mathcal{X}': \|x - x'\| \leq D$,
2. $\forall x_h, x_r, x_t \in \mathcal{X}: \|\psi(x_h, x_r, x_t)\| \leq M$,
3. $\forall x_h, x_r, x_t, x_h', x_r', x_t' \in \mathcal{X}: \|\psi(x_h, x_r, x_t) - \psi(x_h', x_r', x_t')\| \leq K(\|x_h - x_h'\| + \|x_r - x_r'\| + \|x_t - x_t'\|)$.

Then the following holds:

$$\mathbb{E} \left[ \sup_{\mathcal{X}} |NL(X|G) - \mathbb{E}[NL(X)]| \right] \leq \frac{4K|G^*|}{|O|} + M \sqrt{\frac{|G^*|}{|O|^2}} \sqrt{\frac{d \ln (4eD|O|\sqrt{d})}{|O|}}.$$

**7 Pinsker’s Inequality Comes into Play**

In this section we use Pinsker’s inequality to show how the KL-divergence between the data-generating distribution, which generates the observed knowledge graph, and the learned distribution can be used to bound the number of incorrectly predicted triples. In combination with the results from the previous section, this will allow us to bound the expected number of incorrectly predicted triples (this is explained in detail in the next section).

The next theorem bounds the number of incorrectly predicted triples as a function of the KL-divergence of the data-generating and the learned distributions, of the threshold $t$ (used to decide which triple is predicted as positive) and the missingness parameter $\delta$ (the fraction of the triples missing from the ground-truth knowledge graph).

**Theorem 3.** Let $G^*$ be the ground-truth graph on a set of objects $O$ and relations $R$. Let $P$ denote the data-generating distribution induced by $G^*$ with the missingness parameter $\delta$ and let $Q$ denote the learned distribution. Finally, let $t \in (0; 1 - \delta)$ be a threshold. Then the following holds for the cardinality of the set of wrongly predicted triples $F$:

$$|F| \leq \max \left\{ \frac{\ln 2}{2t^2}, \frac{\ln 2}{2(1 - \delta - t)^2} \right\} \cdot KL(P||Q)$$

Note that the bound from this theorem is minimized for the threshold $t := (1 - \delta)/2$. This suggests how to set the threshold if we knew the missingness parameter $\delta$.

**8 Putting Everything So Far Together**

Next we combine the results from the previous sections to obtain a bound on the expected number of incorrectly predicted triples.
**Theorem 4.** Let $\mathcal{G}^*$ be a ground-truth knowledge graph on a set of objects $\mathcal{O}$ and a set of relations $\mathcal{R}$, and let $\mathcal{G}$ be sampled by the data-generating distribution. Let $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\mathcal{H}_X$ ("hypothesis class") denote the set of vector representations $X$, $X \subseteq \mathcal{X}^{(\mathcal{O})+|\mathcal{R}|}$, of the objects from $\mathcal{O}$ and relations from $\mathcal{R}$. Let $\mathcal{K}, \mathcal{M}, \mathcal{D}$ be as in Theorem 2 and $\gamma$ and $\delta$ be as in Theorem 3. Let $X \in \mathcal{H}_X$ be a representation of objects and relations learned by maximizing the log-likelihood $L(X|\mathcal{G})$ and $X^* = \arg \max_{X \in \mathcal{H}_X} E[L(X|\mathcal{G})]$. Then the following holds for the sets of incorrectly predicted triples $\mathcal{F}_X$, obtained using the KG-distribution with $X$:

$$
\mathbb{E}\left[\frac{|\mathcal{F}_X|}{|\mathcal{O}|^2|\mathcal{R}|}\right] \leq \max \left\{ \frac{\ln 2}{t^2 \cdot (1 - \gamma - t)^2}, \frac{\ln 2}{2(t - \gamma)^2}, \frac{4K}{|\mathcal{O}|} \right\} \cdot \left(4K \frac{|\mathcal{G}^*|}{|\mathcal{O}|^2} \cdot \sqrt{\ln (4eD)|\mathcal{O}||\mathcal{D}|} + KL(P||P_{X^*})\right)
$$

where $KL(P||P_{X^*})$ is the Kullback-Leibler divergence of the data-generating distribution $P$ and the best possible representable distribution $P_{X^*}$.

**9 The "Realizable" Case**

In learning theory, one usually speaks of the "realizable case" when the hypothesis class contains the target concept. In general, the data generating distribution cannot be represented exactly using the KG-distribution with vector representations from a given set $\mathcal{X}$ unless we allow $|\mathcal{X}|$ to be infinite (which would be needed to model the triples that have zero probability of being sampled from the data generating distribution because they are not contained in the ground-truth knowledge graph $\mathcal{G}^*$). Therefore we cannot really speak of the "realizable case" in the usual sense. In particular it may be the case that the given scoring function $\psi$ and the given domain of vector representations $\mathcal{X}$ are enough to represent the knowledge graph in the sense that there exists a threshold $t$ such that $(\psi(x_h, x_r, x_t) \geq t) \Leftrightarrow (h, r, t) \in \mathcal{G}^*)$, and yet the data-generating distribution could not be represented by the corresponding KG-distribution exactly. Hence, it makes sense to define realizable case slightly differently for our setting.

**Definition 4 (\(\gamma\)-Realizable case).** For a given data-generating distribution $P$ induced by a ground-truth knowledge graph $\mathcal{G}^*$ with the missingness parameter $\delta$ and a given scoring function $\psi$ and a domain of vector representations $\mathcal{X}$, we say that a KG-distribution $\mathcal{P}_X$ given by a scoring function $\psi$ and vector representations $X$ is $\gamma$-admissible if the following is satisfied: $\exp(\psi(x_h, x_r, x_t))/(1 + \exp(\psi(x_h, x_r, x_t))) \geq 1 - \gamma$ for all $(h, r, t) \in \mathcal{G}^*$ and $1/(1 + \exp(\psi(x_h, x_r, x_t))) \leq \gamma$ for all $(h, r, t) \notin \mathcal{G}^*$, where $x_h, x_r, x_t$ are vector representations given by $X$. We say that a learning problem is $\gamma$-realizable if all vector representations $X \in \mathcal{X}^{(\mathcal{O})+|\mathcal{R}|}$ that maximize the **expected** log-likelihood (i.e. the best vector representations) give rise to $\gamma$-admissible KG-distributions.

Next we present a version of Theorem 3 suitable for $\gamma$-realizable settings.

**Theorem 5.** Let $\mathcal{G}^*$ be the ground-truth graph on a set of objects $\mathcal{O}$ and relations $\mathcal{R}$. Let $P$ denote the data-generating distribution induced by $\mathcal{G}^*$ and let $P_\gamma$ be a $\gamma$-admissible distribution. Let $Q$ denote the learned distribution and let $t \in (\gamma; 1 - \gamma)$ be a threshold. Then the following holds for the cardinality of the set of wrongly predicted triples $\mathcal{F}$:

$$
|\mathcal{F}| \leq \max \left\{ \frac{\ln 2}{2(t - \gamma)^2}, \frac{\ln 2}{2(1 - \gamma - t)^2} \right\} \cdot KL(P_\gamma||Q)
$$

In the next theorem we present a tighter version of Theorem 4 for $\gamma$-realizable settings.

**Theorem 6.** Let everything be as in Theorem 4 but assume a $\gamma$-realizable setting. Then the following holds for the sets of incorrectly predicted triples $\mathcal{F}_X$, obtained using the KG-distribution with $X$:

$$
\mathbb{E}\left[\frac{|\mathcal{F}_X|}{|\mathcal{O}|^2|\mathcal{R}|}\right] \leq \max \left\{ \frac{\ln 2}{(t - \gamma)^2}, \frac{\ln 2}{2(1 - \gamma - t)^2} \right\} \cdot \left(4K \frac{|\mathcal{G}^*|}{|\mathcal{O}|^2} \cdot \sqrt{\ln (4eD)|\mathcal{O}||\mathcal{D}|} + KL(P||P_{X^*})\right)
$$

The next example illustrates that the results from Theorem 6 are asymptotically non-trivial (which is non-obvious as we show below).
Example 2. Let us assume that we have some fixed $H_X$, $K$, $D$ and $d$ and that the $X^* \in H_X$. For simplicity, let us assume that this is a $\gamma$-realizable setting with $\gamma = 0.1$. Let the ground knowledge graphs $\mathcal{G}$ be sparse and satisfy $|\mathcal{G}| = \alpha(|\mathcal{O}|) \cdot |\mathcal{O}|$ where $\alpha$ is an integer-valued function and $\mathcal{O}$ is the set of objects in the knowledge graph. We suppose that the threshold $t$ satisfies $t \in (\gamma, 1 - \gamma)$. Then the expected fraction of triples incorrectly predicted by the learned model from our hypothesis class can be bounded by:

$$C_1 \cdot \sqrt{\alpha(|\mathcal{O}|) \cdot \ln |\mathcal{O}|} / |\mathcal{O}|$$

where $C$ is a constant. Now consider another “alternative prediction method” which always predicts all triples as false. The fraction of triples incorrectly predicted by this method will be:

$$E \left[ |\mathcal{F}| / (|\mathcal{O}|^2 |\mathcal{R}|) \right] \leq C_2 \cdot \alpha(|\mathcal{O}|) / |\mathcal{O}|.$$ 

It follows that whenever $\alpha(|\mathcal{O}|)$ grows larger than $\ln |\mathcal{O}|$, the bounds of the maximum likelihood method, derived in this paper, become tighter than the error bound of this trivial method.

Here, we note that $\alpha(|\mathcal{O}|)$ growing only as a logarithm corresponds to very sparse knowledge graphs (in which prediction of positive examples is likely hard).

10 Existing KGE Methods in This Light

In this section we discuss how existing knowledge graph embedding methods in general fit into our analysis.

Scoring Functions We assumed that predictions are made by comparing $\exp (\psi(x_h, x_r, x_t))$ with a threshold. In existing works, one directly compares the output of the scoring function $\psi(x_h, x_r, x_t)$ with a threshold. This differs only by taking a logarithm and, hence, obviously does not harm applicability of our analysis.

Loss Not all knowledge graph embedding methods optimize log-likelihood, which was assumed throughout our analysis. ComplEx (Trouillon et al. 2016), ConvE (Dettmers et al. 2018) and SimpleE (Kazemi & Poole 2018) are among methods that use (regularized) log-loss. Since log-loss differs from log-likelihood only by a constant factor, our analysis applies to these methods directly (ignoring the regularization). For other methods, log-loss can in principle be also used and it was observed by Trouillon & Nickel (2017) that the margin-based loss functions, used by many knowledge graph embedding methods, are more prone to overfitting compared to log-likelihood. Our results may shed further light on this latter observation.

Lipschitz Continuity Our result requires Lipschitz continuity, Most of the scoring functions $\psi$ used by existing knowledge graph methods are not globally Lipschitz continuous. However, they are Lipschitz continuous if we restrict the domain of the embeddings to a bounded subset of $\mathbb{R}^d$ (this is among others true for the approaches based on matrix or tensor factorization).

Domain of Vector Representations For obtaining the generalization bounds, we needed to assume that the learned vector representations come from a set with a bounded diameter. Existing methods usually do not have such a constraint explicitly but use a regularization term which has a similar effect in practice.

A Numerical Example We finish this section by showing how our results apply to TransE (Bordes et al. [2013]), arguably the simplest embedding method. In TransE, $\psi(x_h, x_r, x_t) = -\|x_h + x_r - x_t\|$. It is not difficult to show that the Lipschitz constant $K = 1$ as follows (using the triangle inequality repeatedly):

$$\|x_h + x_r - x_t\| - \|x_h' + x_r' - x_t'\| = \|x_h - x_h' + x_r - x_r' - x_t + x_t' + x_h' + x_r' - x_t'\|$$

$$-\|x_h + x_r - x_t\| \leq \|x_h - x_h'\| + \|x_r - x_r'\| + \|x_t - x_t'\|.$$ 

Let the dimension of the vector embeddings be $d = 3$. Let us suppose that for all $x \in X$, $\|x\| \leq \ln 10$. Then the remaining parameters from Theorem 3 are $D = \ln 10^2$ and $M = \ln 10^3$. For simplicity, let us assume a $\gamma$-realizable setting with $\gamma = 0.01$. Let $t = 0.5$. Finally suppose that $|\mathcal{G}^*| = |\mathcal{O}| \ln |\mathcal{O}|$. Now we have all parameters needed to apply Theorem 3. We plot the resulting bound for a varying number of objects in Figure 7. In this case the bound becomes non-vacuous for number of objects in the domain $|\mathcal{O}|$ around 1000. Note that, even though the number of dimensions of the vector embeddings is low, the total number of parameters of the model is very high because each object and each relation has its own parameters.
Our work is very close in spirit to the theoretical analyses of learning from positive and unlabeled examples, known as PU-learning (Niu et al., 2016). However, there are several distinctions that do not allow direct application of the existing results to the problem that we analyze in this paper. First, existing works on PU-learning assume the inductive setting whereas our setting is transductive. Second, the work (Niu et al., 2016), which is one of the few papers analyzing generalization bounds for PU-learning, only applies to losses that satisfy \( l(t, +1) + l(t, -1) = 1 \) which is not the case for log-loss. Moreover the losses considered by Niu et al. (2016) are usually not used for knowledge graph embedding in practice. Third, even if the previous two differences were not a problem (although they are), it would still not be obvious how to apply the existing results to the knowledge graph embedding setting in which we deal with interconnected examples; i.e. it is not clear what the attributes of the examples and the hypothesis class should be so that we could fit knowledge graph embedding setting into the standard PU learning setting. Lastly, an advantage of our analysis is that it uses a simple probabilistic model which provides a possible, hopefully intuitive, explanation of why knowledge graph embedding methods work (at least in the standard testing scenarios that use test data generated completely at random).

Another line of works that are related to ours are works on generalization bounds for link prediction, such as (London et al., 2013; Wang et al., 2018). The main differences w.r.t. these works is that (i) they assume availability of both positive and negative examples (whereas we only have positive examples), (ii) the links are predicted based on attributes of nodes and not on the structure of the relations. Finally, our work is also related to works for tensor completion using tensor decomposition, such as (Nickel & Tresp, 2013). However, first, not all knowledge graph embedding methods can be seen as some low-rank tensor decompositions and, second, as far as we know no analysis of generalization bounds of tensor decomposition applies to the PU-learning setting, in which one observes only positive examples (which is the setting studied in our work).

12 Conclusions

In this paper we have derived generalization bounds for the expected number of triples predicted incorrectly by knowledge graph embedding methods. The main assumptions that we used were: (i) facts are missing from the knowledge graph “completely at random” and (ii) the vector representations of objects and relations are learned by maximizing log-likelihood of the model (equivalently by minimizing log-loss). While the “missing completely at random” assumption may seem too simplistic, it is actually how training and test sets are usually constructed, e.g. in (Bordes et al., 2013), and only recently there have been works systematically looking beyond datasets constructed in this way, e.g. (Trouillon et al., 2019). As for the second assumption, it would be interesting to see if one could derive similar bounds for other loss functions that have been used for knowledge graph embedding in the literature, which would then provide a more solid theoretical justification for their use.
REFERENCES


### A Useful Inequalities

We will need the following well-known inequalities.

**Lemma 7** (Pinsker’s inequality (Pinsker [1964])). Let $P$ and $Q$ be distributions. Then

$$\text{KL}(P||Q) \geq \frac{1}{2\ln 2} \|P - Q\|^2.$$ 

**Lemma 8** (McDiarmid’s Inequality (McDiarmid [1989])). Let $X_1, \ldots, X_n$ be independent random variables with values from a set $\mathcal{X}$ and let $f : \mathcal{X}^n \rightarrow \mathbb{R}$. If, for all $i \in \{1, \ldots, n\}$, and for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$, the function $f$ satisfies $|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$ then:

$$P[|f(X_1, \ldots, X_n) - \mathbb{E}[f]| \geq \varepsilon] \leq 2 \exp \left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

### B Proof of Theorem \[ \text{[1]} \]

In this section we prove Theorem \[ \text{[1]} \]

**Theorem 1.** Let $G^*$ be the ground-truth knowledge graph. Let $G$ be sampled from the data-generating distribution induced by $G^*$ and $\mathcal{X}$ be fixed vector representations. Then

$$P[|L(\mathcal{X}|G) - \mathbb{E}[L(\mathcal{X})]| \geq \varepsilon] \leq 2 \exp \left(-\frac{2\varepsilon^2}{M^2\|G^*\|}\right),$$

where $M \geq \sup_{h,t \in \mathcal{X}_v, r \in \mathcal{R}_v} |\psi(h, r, x_t)|$.

**Proof.** Let $F := |G^*|$. Let $U_1, \ldots, U_F$ be Bernoulli random variables such that $P[U_i = 1] = 1 - \delta$. Let us denote by $\tau_i$ the $i$-th tuple in $G^*$. Let $g : \{0, 1\}^F \rightarrow 2^{G^*}$ be defined as $g(u_1, \ldots, u_F) = \{\tau_i | u_i = 1\}$, i.e. $g(u_1, \ldots, u_F)$ contains all $\tau_i$’s for which $u_i = 1$. It is not difficult to see that $G$ and $g(U_1, \ldots, U_F)$ have the same distribution. Hence we can bound the probability $P[|L(\mathcal{X}|g(U_1, \ldots, U_F)) - \mathbb{E}[L(\mathcal{X})]| \geq \varepsilon]$ instead of directly $P[|L(\mathcal{X}|G) - \mathbb{E}[L(\mathcal{X})]| \geq \varepsilon]$.

It follows from the assumptions of the theorem that for any $(u_1, \ldots, u_F) \in \{0, 1\}^F$ the following must hold:

$$|L(\mathcal{X}|g(u_1, \ldots, u_1, \ldots, u_F)) - L(\mathcal{X}|g(u_1, \ldots, u'_1, \ldots, u_F))|$$

$$= \left| \sum_{(h,r,t) \in g(u_1, \ldots, u_1, \ldots, u_F)} \psi(h, r, x_t) - \ln Z - \sum_{(h,r,t) \in g(u_1, \ldots, u'_1, \ldots, u_F)} \psi(h, r, x_t) + \ln Z \right|$$

$$\leq |\psi(h, r, x_t)| \leq M,$
where \( x_h, x_r \) and \( x_t \) denote vector representations of \( h, r \) and \( t \), respectively.

Hence, we can use McDiarmid’s inequality to obtain
\[
P \left[ |L(X|g(U_1, \ldots, U_F)) - E[L(X_\circ)]| \geq \varepsilon \right] \leq 2 \exp \left( -\frac{2\varepsilon^2}{M^2} \right)
\]
from which the statement of the theorem follows directly. \( \square \)

### C Proof of Theorem 2

In this section we give a proof of Theorem 2. First, we prove several lemmas that we need to prove this theorem.

**Lemma 9.** Let \( \tilde{M} > 0 \) be such that for every object or relation \( o \) it holds \( \|x_o - x'_o\| \leq \tilde{M} \) where \( x_o \) and \( x'_o \) are the vector representations of \( o \) in \( X \) and \( X' \), respectively. If \( \psi \) satisfies the assumptions of Theorem 2 then the following holds:
\[
|L(X_o, X_r, g) - L(X'_o, X'_r, g)| \leq 2K|O|^2|R|\tilde{M}.
\]

**Proof.** Let us first introduce some notation. Let \( V := |O| \) and \( R = |R| \). Whenever we write \( x_h, x_r \) and \( x_t \), it denotes the vector representations of \( h, r \) and \( t \) from \( X \), whereas when we write \( x'_h, x'_r \) and \( x'_t \), it denotes the representations from \( X' \). We have
\[
|L(X_o, X_r, g) - L(X'_o, X'_r, g)| = \left| \sum_{(h,r,t) \in G} (\psi(x_h, x_r, x_t) - \psi(x'_h, x'_r, x'_t)) \right| \leq \left| \sum_{(h,r,t) \in G} (\psi(x_h, x_r, x_t) - \psi(x'_h, x'_r, x'_t)) \right| + \ln \frac{Z}{Z'}.
\]

Next we bound the terms on the r.h.s. of the above inequality separately. We start by obtaining the bound for the first term:
\[
\left| \sum_{(h,r,t) \in G} (\psi(x_h, x_r, x_t) - \psi(x'_h, x'_r, x'_t)) \right| \leq KV^2 R \cdot (\|x_h - x'_h\| + \|x_r - x'_r\| + \|x_t - x'_t\|) \leq KV^2 R \tilde{M}.
\]

It remains to bound the second term. We have:
\[
\frac{Z}{Z'} = \frac{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in G} e^{\psi(x_h, x_r, x_t)}}{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in \tilde{G}} e^{\psi(x_h, x_r, x_t)}} = \frac{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in G} e^{(\psi(x_h, x_r, x_t) - \psi(x'_h, x'_r, x'_t)) + (\psi(x'_h, x'_r, x'_t))}}{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in \tilde{G}} e^{\psi(x'_h, x'_r, x'_t)}} \leq e^{K \cdot V^2 R \tilde{M}} \frac{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in G} e^{\psi(x'_h, x'_r, x'_t)}}{\sum_{G \subseteq \Omega} \prod_{(h,r,t) \in \tilde{G}} e^{\psi(x'_h, x'_r, x'_t)}} = e^{K \cdot V^2 R \tilde{M}}
\]

where the last inequality follows from the inequality \( K (\|x_h - x'_h\| + \|x_r - x'_r\| + \|x_t - x'_t\|) \leq K \tilde{M} \). Combining (1) and (2) we obtain the inequality that we needed to prove. \( \square \)

**Lemma 10.** Let \( G^* \) be a ground-truth knowledge graph on a set of objects \( \mathcal{O} \) and a set of relations \( \mathcal{R} \) and let \( G \) be sampled by the data-generating distribution. Let \( \mathcal{H} \) (“hypothesis class”) be a finite set of vector representations of the objects from \( \mathcal{O} \) and relations from \( \mathcal{R} \). Then the following inequality holds
\[
\mathbb{E} \left[ \sup_{X \in \mathcal{H}} |L(X|g) - \mathbb{E}[L(X_\circ)]| \right] \leq M \sqrt{\frac{|G^*| \ln (2e|\mathcal{H}|)}{2}}
\]
where \( M \geq \sup_{h,t \in \mathcal{O}, r \in \mathcal{R}} |\psi(x_h, x_r, x_t)|. \)
Proof. We denote $W := \sup_{X \in \mathcal{H}} \left| L(X|G) - E[L(X|x)] \right|$ and $B := 2/\left( |\mathcal{G}^*| \cdot M^2 \right)$. Then, using the union bound and Theorem[1] we have: $P \left[ W \geq \varepsilon \right] \leq 2 |\mathcal{H}| \cdot \exp \left( -\varepsilon^2 \cdot B \right)$. Next, we have

$$\mathbb{E}[W] \leq \sqrt{\mathbb{E}[W^2]} = \sqrt{\int_0^\infty P[W^2 \geq t] dt} = \sqrt{\int_0^\infty P[W \geq \sqrt{t}] dt} \leq \sqrt{u + \int_u^\infty P[W \geq \sqrt{t}] dt} \leq \sqrt{u + \int_u^\infty 2|\mathcal{H}| \cdot \exp (-t \cdot B) dt} = \sqrt{u + \frac{2|\mathcal{H}|}{B} \cdot \exp (-u \cdot B)}$$

The above expression is minimized for $u := \ln \left( \frac{2|\mathcal{H}|}{B} \right)$ (note that $u$ may be arbitrary), yielding the bound:

$$\sqrt{\frac{\ln (2e|\mathcal{H}|)}{B}} = M \sqrt{\frac{|\mathcal{G}^*| \ln (2e|\mathcal{H}|)}{2}}$$

Lemma 11 (Covering number, e.g., [Shalev-Shwartz & Ben-David, 2014]). Let $S$ be a subset of $\mathbb{R}^d$ of diameter at most $k$ (i.e. for any $x_1, x_2 \in S$, $\|x_1 - x_2\| \leq k$). Then $S$ can be covered by $(2k\sqrt{d}/\rho)^d$ balls of radius $\rho$.

Now we are ready to prove Theorem 2.

Theorem 2. Let $\mathcal{G}^*$ be a ground-truth knowledge graph on a set of objects $\mathcal{O}$ and a set of relations $\mathcal{R}$ and let $G$ be sampled by the data-generating distribution. Let $X \subseteq \mathbb{R}^d$. Finally, let $\mathcal{H}_X$ ("hypothesis class") denote the set of vector representations $X, X \subseteq X^{|\mathcal{O}|+|\mathcal{R}|}$, of the objects from $\mathcal{O}$ and relations from $\mathcal{R}$. Let $D$, $M$ and $K$ be real numbers satisfying:

1. $\forall x, x' \in X: \|x - x'\| \leq D$,
2. $\forall x_i, x_j, x_k \in X: |\psi(x_i, x_j, x_k)| \leq M$,
3. $\forall x_i, x_j, x_k, x_i', x_j', x_k' \in X: \|\psi(x_i, x_j, x_k) - \psi(x_i', x_j', x_k')\| \leq K(\|x_i - x_i'\| + \|x_j - x_j'\| + \|x_k - x_k'\|)$.

Then the following holds:

$$\mathbb{E} \left[ \sup_{X \in \mathcal{H}_X} \left| NL(X|G) - E[NL(X|x)] \right| \right] \leq \frac{4K}{|\mathcal{O}|} + M \sqrt{\frac{|\mathcal{G}^*|}{|\mathcal{O}|^2}} \sqrt{\frac{d \ln (4eD|\mathcal{O}|\sqrt{d})}{|\mathcal{O}|}}$$

Proof. Let us denote $V := |\mathcal{O}|$ and $R := |\mathcal{R}|$. Let $B$ be a finite set of vectors of dimension $d$ such that if we place centers of balls of radius $\rho$ in all these points then these balls will cover the set $X$. Using Lemma[11] we know that it is possible to find such a set $B$ satisfying

$$|B| \leq \left( 2D\sqrt{d}/\rho \right)^d$$

(3)

because the diameter of the set $X$ is at most $D$ (due to the assumptions of the theorem) and the dimension of the vectors in $X$ is $d$. Next we assume that we are only searching for a maximum-likelihood solution $X$ consisting of vectors from $B$ so that we could use Lemma[10] which expects a finite hypothesis set $\mathcal{H}$. Even though we have so far been thinking of the vector representations $X$ as lookup tables, it will be more convenient here to think of them as lists of vectors. Specifically, we assume that the objects and relations are ordered arbitrarily. The hypothesis set will then be represented as $\mathcal{H}_B = B^{(|\mathcal{O}|+|\mathcal{R}|)}$. Hence, using (3), we obtain $|\mathcal{H}_B| \leq \left( 2D\sqrt{d}/\rho \right)^{d(|\mathcal{O}|+|\mathcal{R}|)}$. Then from Lemma[10] we have:

$$\mathbb{E} \left[ \sup_{X \in \mathcal{H}_B} \left| L(X|G) - E[L(X|x)] \right| \right] \leq M \sqrt{\frac{|\mathcal{G}^*| \ln (2e|\mathcal{H}_B|)}{2}}$$

(4)

Using the bound on $|\mathcal{H}_B|$, we can bound the r.h.s. by:

$$M \sqrt{\frac{|\mathcal{G}^*| d(V + R) \ln \left( \frac{4eD\sqrt{d}}{\rho} \right)}{2}}$$

(5)
For fixed vector representations $\mathcal{X}'$ of the objects and relations, let $\mathcal{H}(\mathcal{X}')$ denote the set of vector representations $\mathcal{X}$ which satisfy that for every object or relation $o$: $|x_o - x'_o| \leq \rho$ where $x_o$ and $x'_o$ are vector representations of $o$ from $\mathcal{X}$ and $\mathcal{X}'$, respectively. Now, optimizing over the hypothesis set $\mathcal{H}_X$ of all possible vector representations from $\mathcal{X}$ instead of just the vectors from the set $\mathcal{H}_B$, would yield the next bound:

$$
E \left[ \sup_{\mathcal{X} \in \mathcal{H}_X} |L(\mathcal{X}|\mathcal{G}) - E[L(\mathcal{X}|.)]| \right] 
= E \left[ \sup_{\mathcal{X}' \in \mathcal{H}_B} \sup_{\mathcal{X} \in \mathcal{H}(\mathcal{X}')} |L(\mathcal{X}|\mathcal{G}) - E[L(\mathcal{X}'|.)] + E[L(\mathcal{X}'|\mathcal{G})] - E[L(\mathcal{X}|.)]| \right] 
\leq E \left[ \sup_{\mathcal{X}' \in \mathcal{H}_B} \sup_{\mathcal{X} \in \mathcal{H}(\mathcal{X}')} |L(\mathcal{X}|\mathcal{G}) - L(\mathcal{X}'|\mathcal{G}) + L(\mathcal{X}'|\mathcal{G}) - E[L(\mathcal{X}'|.)]| \right] 
+ E \left[ \sup_{\mathcal{X}' \in \mathcal{H}_B} \sup_{\mathcal{X} \in \mathcal{H}(\mathcal{X}')} |E[L(\mathcal{X}'|.)] - E[L(\mathcal{X}|.)]| \right]
$$

Next, to bound the first and third term, we use Lemma 9 and, to bound the second term, we use Equations (4) and (5). After setting $\rho := \frac{1}{4}$ and after simplifying, this gives us the following bound:

$$
E \left[ \sup_{\mathcal{X} \in \mathcal{H}_X} |L(\mathcal{X}|\mathcal{G}) - E[L(\mathcal{X}|.)]| \right] \leq 4KRV + M \sqrt{\frac{|G^*|dR(V + R) \ln (4eDV\sqrt{d})}{2}} 
\leq 4KRV + MR \sqrt{|G^*| \cdot Vd \ln (4eDV\sqrt{d})}.
$$

For the expected error of normalized log-likelihood we then obtain

$$
E \left[ \sup_{\mathcal{X} \in \mathcal{H}_X} |NL(\mathcal{X}|\mathcal{G}) - E[NL(\mathcal{X}|.)]| \right] \leq 4K \cdot \sqrt{\frac{|G^*|}{V^2}} \cdot \sqrt{\frac{d \ln (4eDV\sqrt{d})}{V}}
$$

which finishes the proof of the theorem.

\[\square\]

D PROOF OF THEOREM 3

In this section we prove Theorem 3.

**Theorem 3.** Let $G^*$ be the ground-truth graph on a set of objects $O$ and relations $R$. Let $P$ denote the data-generating distribution induced by $G^*$ with the missingness parameter $\delta$ and let $Q$ denote the learned distribution. Finally, let $t \in (0; 1 - \delta)$ be a threshold. Then the following holds for the cardinality of the set of wrongly predicted triples $\mathcal{F}$:

$$
|\mathcal{F}| \leq \max \left\{ \frac{\ln 2}{2t^2}, \frac{\ln 2}{2(1 - \delta - t)^2} \right\} \cdot KL(P||Q)
$$

In the proof of the theorem, we will use the following simple lemma (which can be verified by computing the discriminant of the respective quadratic equation).

**Lemma 12.** The following inequality holds for all $x, y \in [0; 1]^2$ and $\alpha \in (0; \infty)$:

$$
\alpha(x - y)^2 + \frac{1}{4\alpha} \geq |x - y|.
$$
Proof of Theorem\cite{12} We assume that all possible triples $(h, r, t) \in O \times R \times O$ are ordered (arbitrarily) and that $\tau_i$ denotes the $i$-th triple from this set and $\mathcal{I} = \{1, 2, \ldots, |O|^2 \cdot |R|\}$. Let $\mathcal{G}$ be sampled from the data-generating distribution and $\mathcal{G}'$ be sampled from the learned model (i.e. the KG-distribution). We define binary random variables $B_i := 1(\tau_i \in \mathcal{G})$ and $B_i' := 1(\tau_i \in \mathcal{G}')$. Hence the complete vectors $B = (B_1, \ldots, B_n)$ and $B' = (B'_1, \ldots, B'_n)$ represent the knowledge graphs sampled from the data-generating and modelled distributions, respectively. We denote by $P_B$ and $P_B'$ the distributions of $B$ and $B'$. Due to the independence properties of the two distributions, the following holds: $KL(P_B || P_{B'}) = \sum_{i \in \mathcal{I}} KL(P_B(i) || P_{B'}(i))$ where $P_{B_i}$ and $P_{B'_i}$ are the distributions of $B_i$ and $B'_i$, respectively.

Next we bound the error that is made when we classify all triples $\tau_i$ with $P[B_i'] = 1 > t$ as true (i.e. belonging to the knowledge graph) and the rest as false. We may notice that when misclassifying a triple $\tau_i$, we must either have

$$P[B_i = 1] = 0 \text{ and } P[B'_i = 1] > t$$

(“false positive”) or

$$P[B_i = 1] = 1 - \delta \text{ and } P[B'_i = 1] < t$$

(“false negative”). Hence, if we denote by $\mathcal{FP}$ the set of triples $\tau_i$ for which we have false positive errors, we must have for all $\alpha > 0$

$$t|\mathcal{FP}| \leq \sum_{\tau_i \in \mathcal{FP}} P[B'_i = 1] \leq \sum_{\tau_i \in \mathcal{FP}} (|P[B_i = 1] - P[B'_i = 1]|) \leq \sum_{\tau_i \in \mathcal{FP}} \left( \frac{1}{4\alpha} + \alpha |P[B_i = 1] - P[B'_i = 1]|^2 \right)$$

where the last inequality follows from Lemma\cite{12}. It follows that

$$|\mathcal{FP}| \cdot \left( t - \frac{1}{4\alpha} \right) \leq \sum_{\tau_i \in \mathcal{FP}} \alpha |P[B_i = 1] - P[B'_i = 1]|^2.$$  \hspace{1cm} (6)

Letting $\alpha = 1/(2t)$ yields the bound

$$|\mathcal{FP}| \leq \frac{1}{t} \cdot \sum_{\tau_i \in \mathcal{FP}} |P[B_i = 1] - P[B'_i = 1]|^2.$$  \hspace{1cm} (6)

Similarly for false negatives, if we denote by $\mathcal{FN}$ the set of triples for which we have false negative errors, we must have for all $\alpha' > 0$:

$$|\mathcal{FN}| \cdot (1 - \delta - t) \leq \sum_{\tau_i \in \mathcal{FN}} |P[B_i = 1] - P[B'_i = 1]|$$

$$\leq \sum_{\tau_i \in \mathcal{FN}} \left( \frac{1}{4\alpha'} + \alpha' |P[B_i = 1] - P[B'_i = 1]|^2 \right).$$

Hence, we also have:

$$|\mathcal{FN}| \cdot \left( \frac{4\alpha' - 1}{4\alpha'} - \delta - t \right) \leq \alpha' \sum_{\tau_i \in \mathcal{FN}} |P[B_i = 1] - P[B'_i = 1]|^2$$

Letting $\alpha' = 1/(2 - 2\delta - 2t)$ gives us

$$|\mathcal{FN}| \leq \frac{1}{(1 - \delta - t)^2} \cdot \sum_{\tau_i \in \mathcal{FN}} |P[B_i = 1] - P[B'_i = 1]|^2.$$  \hspace{1cm} (7)
Next we combine (6) and (7) and apply Pinsker’s inequality:

\[
|\mathcal{F}_P| + |\mathcal{F}_N| \leq \frac{1}{t^2} \cdot \sum_{\tau_i \in \mathcal{F}_P} |P[B_i = 1] - P[B'_i = 1]|^2 \\
+ \frac{1}{(1-\delta-t)^2} \cdot \sum_{\tau_i \in \mathcal{F}_N} |P[B_i = 1] - P[B'_i = 1]|^2 \\
= \frac{1}{t^2} \cdot \sum_{\tau_i \in \mathcal{F}_P} \frac{1}{4} ||P_{B_i} - P_{B'_i}||^2_1 + \frac{1}{(1-\delta-t)^2} \cdot \sum_{\tau_i \in \mathcal{F}_N} \frac{1}{4} ||P_{B_i} - P_{B'_i}||^2_1 \\
\leq \frac{1}{4} \max \left\{ \frac{\ln 2}{t^2}, \frac{\ln 2}{(1-\delta-t)^2} \right\} \cdot \sum_{\tau_i \in I} ||P_{B_i} - P_{B'_i}||^2_1 \\
\leq \max \left\{ \frac{\ln 2}{2t^2}, \frac{\ln 2}{2(1-\delta-t)^2} \right\} \cdot KL(P_B \Vert P_{B'}). \quad (8)
\]

This finishes the proof of this theorem.

\[\square\]

### E Proof of Theorem 4

**Theorem 4** Let \( \mathcal{G}^\ast \) be a ground-truth knowledge graph on a set of objects \( \mathcal{O} \) and a set of relations \( \mathcal{R} \) and let \( \mathcal{G} \) be sampled by the data-generating distribution. Let \( \mathcal{X} \subseteq \mathbb{R}^d \). Let \( \mathcal{H}_\mathcal{X} \) (“hypothesis class”) denote the set of vector representations \( \mathcal{X} \subseteq \mathcal{X}^{(|\mathcal{O}|+|\mathcal{R}|)} \) of the objects from \( \mathcal{O} \) and relations from \( \mathcal{R} \). Let \( K, M, D, t \) be as in Theorem 2 and \( \delta \) be as in Theorem 3. Let \( \mathcal{X} \in \mathcal{H}_\mathcal{X} \) be a representation of objects and relations learned by maximizing the log-likelihood \( L(\mathcal{X} \mid \mathcal{G}) \) and \( \mathcal{X}^\ast = \arg \max_{\mathcal{X} \in \mathcal{H}_\mathcal{X}} E[L(\mathcal{X} \mid \mathcal{G})] \). Then the following holds for the sets of incorrectly predicted triples \( \mathcal{F}_\mathcal{X} \), obtained using the KG-distribution with \( \mathcal{X} \):

\[
E \left[ \frac{|\mathcal{F}_\mathcal{X}|}{|\mathcal{O}|^2 |\mathcal{R}|} \right] \leq \max \left\{ \frac{\ln 2}{t^2}, \frac{\ln 2}{(1-\delta-t)^2} \right\} \cdot \left( \frac{4K}{|\mathcal{O}|} \\
+ M \sqrt{\frac{|\mathcal{G}^\ast|}{|\mathcal{O}|^2}} \cdot \sqrt{\frac{d \ln (4eD|\mathcal{O}| \sqrt{d})}{|\mathcal{O}|^2 |\mathcal{R}|^2}} + \frac{KL(P \Vert P_{\mathcal{X}^\ast})}{|\mathcal{O}|^2 |\mathcal{R}|} \right)
\]

where \( KL(P \Vert P_{\mathcal{X}^\ast}) \) is the Kullback-Leibler divergence of the data-generating distribution \( P \) and the best possible distribution \( P_{\mathcal{X}^\ast} \).

**Proof.** Let \( \mathcal{G} \) denote the knowledge graph sampled from the data-generating distribution and \( P_{\mathcal{X}} \) the learned distribution. For a fixed \( \mathcal{X}_{fixed} \), the following equality holds:

\[
KL(P \Vert P_{\mathcal{X}_{fixed}}) = H(P) + E_P[L(\mathcal{X}_{fixed} \mid \mathcal{G})].
\]

Then we also have

\[
KL(P \Vert P_{\mathcal{X}_{fixed}}) - KL(P \Vert P_{\mathcal{X}^\ast}) = E_P[L(\mathcal{X}_{fixed} \mid \mathcal{G})] - E_P[L(\mathcal{X}^\ast \mid \mathcal{G})].
\]

Next we can define \( g(\mathcal{G}') := \arg \max_{\mathcal{X} \in \mathcal{H}_\mathcal{X}} E[L(\mathcal{X} \mid \mathcal{G}') \mid \mathcal{G}'] \), substitute \( g(\mathcal{G}') \) for \( \mathcal{X}_{fixed} \) and take the expectation over graphs \( \mathcal{G}' \), which are assumed to be sampled from the data-generating distribution. This yields:

\[
E_{\mathcal{G}' \sim P}[KL(P \Vert P_{g(\mathcal{G}')})] - KL(P \Vert P_{\mathcal{X}^\ast}) = E_{\mathcal{G}' \sim P}[E_{\mathcal{X} \sim P}[L(\mathcal{X} \mid \mathcal{G}') \mid \mathcal{G}')] - E_{\mathcal{X} \sim P}[L(\mathcal{X}^\ast \mid \mathcal{G}^\ast)].
\]

Next we can use Lemma 8.2 from [Devroye et al., 2013] to bound the r.h.s.:

\[
E_{\mathcal{G}' \sim P}[L(\mathcal{X} \mid \mathcal{G}')] - E_{\mathcal{X} \sim P}[L(\mathcal{X}^\ast \mid \mathcal{G}^\ast)] \leq 2E_{\mathcal{G}' \sim P} \left[ \sup_{\mathcal{X} \in \mathcal{H}_\mathcal{X}} |L(\mathcal{X} \mid \mathcal{G}') - L(\mathcal{X}^\ast \mid \mathcal{G}^\ast)| \right].
\]

Finally, we use the bound from Theorem 3 to bound the r.h.s. of this inequality. Combining this with Theorem 4 finishes the proof.

\[\square\]

### F Proof of Theorem 5

**Theorem 5** Let \( \mathcal{G}^\ast \) be the ground-truth graph on a set of objects \( \mathcal{O} \) and relations \( \mathcal{R} \). Let \( P \) denote the data-generating distribution induced by \( \mathcal{G}^\ast \) and let \( P' \) be a \( \gamma \)-admissible distribution. Let \( Q\)
denote the learned distribution and let $t \in (\gamma; 1 - \gamma)$ be a threshold. Then the following holds for the cardinality of the set of wrongly predicted triples $\mathcal{F}$:

$$|\mathcal{F}| \leq \max \left\{ \frac{\ln 2}{2(t - \gamma)^2}, \frac{\ln 2}{2(1 - \gamma - t)^2} \right\} \cdot KL(P_\gamma || Q).$$

**Proof.** The proof of this theorem is very similar to the proof of Theorem 3. We highlight the most important differences in bold.

We assume that all possible triples $(h, r, t) \in \mathcal{O} \times \mathcal{R} \times \mathcal{O}$ are ordered (arbitrarily) and that $\tau_i$ denotes the $i$-th triple from this set and $\mathcal{I} = \{1, 2, \ldots, |\mathcal{O}|^2 \cdot |\mathcal{R}|\}$. Let $\mathcal{G}$ be sampled from the distribution $P_\gamma$ and $\mathcal{G}'$ be sampled from the learned model (i.e. the KG-distribution). We define binary random variables $B_i := 1(\tau_i \in \mathcal{G})$ and $B_i' := 1(\tau_i \in \mathcal{G}')$. Hence the complete vectors $\mathbf{B} = (B_1, \ldots, B_n)$ and $\mathbf{B}' = (B_1', \ldots, B_n')$ represent the knowledge graphs sampled from the distribution $P_\gamma$ and the modelled distribution, respectively. We denote by $P_B$ and $P'_B$ the distributions of $\mathbf{B}$ and $\mathbf{B}'$. Due to the independence properties of the two distributions (after all, $P_\gamma$ is still a KG-distribution and therefore inherits its independence structure), the following holds:

$$KL(P_B || P'_B) = \sum_{i \in \mathcal{I}} KL(P_{B_i} || P'_{B_i})$$

where $P_{B_i}$ and $P'_{B_i}$ are the distributions of $B_i$ and $B_i'$, respectively.

Next we bound the error that is made when we classify all triples $\tau_i$ with $P[B_i = 1] > t$ as true (i.e. belonging to the knowledge graph) and the rest as false. We may notice that when misclassifying a triple $\tau_i$, we must either have

$$P[B_i = 1] \leq \gamma \text{ and } P[B_i' = 1] > t$$

(“false positive”) or

$$P[B_i = 1] \geq 1 - \gamma \text{ and } P[B_i' = 1] < t$$

(“false negative”).

The rest of differences w.r.t. the proof of Theorem 3 follows from the differences highlighted so far in a straightforward way (details follow below).

Hence, if we denote by $\mathcal{FP}$ the set of triples $\tau_i$ for which we have false positive errors, we must have for all $\alpha > 0$

$$t |\mathcal{FP}| \leq \sum_{\tau_i \in \mathcal{FP}} P[B_i' = 1] \leq \sum_{\tau_i \in \mathcal{FP}} (\gamma + |P[B_i = 1] - P[B_i' = 1]|)$$

$$\leq \sum_{\tau_i \in \mathcal{FP}} \left( \gamma + \frac{1}{4\alpha} + \alpha |P[B_i = 1] - P[B_i' = 1]|^2 \right)$$

where the last inequality follows from Lemma 12. It follows that

$$|\mathcal{FP}| \cdot \left( t - \gamma - \frac{1}{4\alpha} \right) \leq \sum_{\tau_i \in \mathcal{FP}} \alpha |P[B_i = 1] - P[B_i' = 1]|^2.$$

Letting $1/(2t - 2\gamma)$ yields the bound

$$|\mathcal{FP}| \leq \frac{1}{(t - \gamma)^2} \cdot \sum_{\tau_i \in \mathcal{FP}} |P[B_i = 1] - P[B_i' = 1]|^2. \quad (9)$$

Similarly for false negatives, if we denote by $\mathcal{FN}$ the set of triples for which we have false negative errors, we must have for all $\alpha' > 0$:

$$|\mathcal{FN}| \cdot (1 - \gamma - t) \leq \sum_{\tau_i \in \mathcal{FN}} |P[B_i = 1] - P[B_i' = 1]|$$

$$\leq \sum_{\tau_i \in \mathcal{FN}} \left( \frac{1}{4\alpha'} + \alpha' |P[B_i = 1] - P[B_i' = 1]|^2 \right).$$

Under review as a conference paper at ICLR 2020
Hence, we also have:

$$|\mathcal{F}_N| \cdot \left( \frac{4\alpha' - 1}{4\alpha'} - \gamma - t \right) \leq \alpha' \sum_{\tau_i \in \mathcal{F}_N} |P[B_i = 1] - P[B'_i = 1]|^2$$

Letting $\alpha' = 1/(2 - 2\gamma - 2t)$ gives us

$$|\mathcal{F}_N| \leq \frac{1}{(1 - \gamma - t)^2} \cdot \sum_{\tau_i \in \mathcal{F}_N} |P[B_i = 1] - P[B'_i = 1]|^2.$$  (10)

Next we combine (9) and (10) and apply Pinsker’s inequality:

$$|\mathcal{F}_P| + |\mathcal{F}_N| \leq \frac{1}{(t - \gamma)^2} \cdot \sum_{\tau_i \in \mathcal{F}_P} |P[B_i = 1] - P[B'_i = 1]|^2$$

$$+ \frac{1}{(1 - \gamma - t)^2} \cdot \sum_{\tau_i \in \mathcal{F}_N} |P[B_i = 1] - P[B'_i = 1]|^2$$

$$= \frac{1}{(t - \gamma)^2} \cdot \sum_{\tau_i \in \mathcal{F}_P} \frac{1}{4} \|P_{B_i} - P_{B'_i}\|^2 + \frac{1}{(1 - \gamma - t)^2} \cdot \sum_{\tau_i \in \mathcal{F}_N} \frac{1}{4} \|P_{B_i} - P_{B'_i}\|^2$$

$$\leq \frac{1}{4} \max \left\{ \frac{1}{(t - \gamma)^2}, \frac{1}{(1 - \gamma - t)^2} \right\} \cdot \sum_{\tau_i \in \mathcal{F}_P} \|P_{B_i} - P_{B'_i}\|^2$$

$$\leq \max \left\{ \frac{\ln 2}{2(t - \gamma)^2}, \frac{\ln 2}{2(1 - \gamma - t)^2} \right\} \cdot KL(P_B || P_{B'}) \cdot (\sum_{\tau_i \in \mathcal{F}_P} \|P_{B_i} - P_{B'_i}\|^2).$$  (11)

This finishes the proof of this theorem. □

G PROOF OF THEOREM 6

**Theorem 6.** Let everything be as in Theorem 4 but assume a $\gamma$-realizable setting. Then the following holds for the sets of incorrectly predicted triples $\mathcal{F}_X$, obtained using the KG-distribution with $X$:

$$E \left[ \frac{\left| \mathcal{F}_X \right|}{|O|^2 |R|} \right] \leq \max \left\{ \frac{\ln 2}{(t - \gamma)^2}, \frac{\ln 2}{(1 - t - \gamma)^2} \right\} \cdot \left( 4K |O|^2 + M \sqrt{\frac{|G|^2}{|O|^2} \ln \left( \frac{4eD|O|}{|O|} \right) \sqrt{d}} \right)$$

**Proof.** The proof of this theorem is practically identical to the proof of Theorem 4. The only difference is that in this proof we invoke Theorem 5 instead of Theorem 4. □