
Knowledge Graph Completion by Intermediate Variables Regularization

Changyi Xiao, Yixin Cao*

School of Computer Science, Fudan University
changyi_xiao@fudan.edu.cn, caoyixin2011@gmail.com

Abstract

Knowledge graph completion (KGC) can be framed as a 3-order binary tensor completion task. Tensor decomposition-based (TDB) models have demonstrated strong performance in KGC. In this paper, we provide a summary of existing TDB models and derive a general form for them, serving as a foundation for further exploration of TDB models. Despite the expressiveness of TDB models, they are prone to overfitting. Existing regularization methods merely minimize the norms of embeddings to regularize the model, leading to suboptimal performance. Therefore, we propose a novel regularization method for TDB models that addresses this limitation. The regularization is applicable to most TDB models and ensures tractable computation. Our method minimizes the norms of intermediate variables involved in the different ways of computing the predicted tensor. To support our regularization method, we provide a theoretical analysis that proves its effect in promoting low trace norm of the predicted tensor to reduce overfitting. Finally, we conduct experiments to verify the effectiveness of our regularization technique as well as the reliability of our theoretical analysis. The code is available at <https://github.com/changyi7231/IVR>.

1 Introduction

A knowledge graph (KG) can be represented as a 3rd-order binary tensor, in which each entry corresponds to a triplet of the form $(head\ entity, relation, tail\ entity)$. A value of 1 denotes a known true triplet, while 0 denotes a false triplet. Despite containing a large number of known triplets, KGs are often incomplete, with many triplets missing. Consequently, the 3rd-order tensors representing the KGs are incomplete. The objective of knowledge graph completion (KGC) is to infer the true or false values of the missing triplets based on the known ones, i.e., to predict which of the missing entries in the tensor are 1 or 0.

A number of models have been proposed for KGC, which can be classified into translation-based models, tensor decomposition-based (TDB) models and neural networks models [Zhang et al., 2021]. We only focus on TDB models in this paper due to their wide applicability and great performance [Lacroix et al., 2018, Zhang et al., 2020]. TDB models can be broadly categorized into two groups: CANDECOMP/PARAFAC (CP) decomposition-based models, including CP [Lacroix et al., 2018], DistMult [Yang et al., 2014] and ComplEx [Trouillon et al., 2017], and Tucker decomposition-based models, including Simple [Kazemi and Poole, 2018], ANALOGY [Liu et al., 2017], QuatE [Zhang et al., 2019] and TuckER [Balažević et al., 2019]. To provide a thorough understanding of TDB models, we present a summary of existing models and derive a general form that unifies them, which provides a fundamental basis for further exploration of TDB models.

*Corresponding author.

TDB models have been proven to be theoretically fully expressive [Trouillon et al., 2017, Kazemi and Poole, 2018, Balažević et al., 2019], implying they can represent any real-valued tensor. However, in practice, TDB models frequently fall prey to severe overfitting. To counteract this issue, various regularization techniques have been employed in KGC. One commonly used technique is the squared Frobenius norm regularization [Nickel et al., 2011, Yang et al., 2014, Trouillon et al., 2017]. Lacroix et al. [2018] proposed another regularization, N3 norm regularization, based on the tensor nuclear p -norm, which outperforms the squared Frobenius norm in terms of performance. Additionally, Zhang et al. [2020] introduced DURA, a regularization technique based on the duality of TDB models and distance-based models that results in significant improvements on benchmark datasets. Nevertheless, N3 and DURA rely on CP decomposition, limiting their applicability to CP [Lacroix et al., 2018] and ComplEx [Trouillon et al., 2017]. As a result, there is a pressing need for a regularization technique that is widely applicable and can effectively alleviate the overfitting issue.

In this paper, we introduce a novel regularization method for KGC to improve the performance of TDB models. Our regularization focuses on preventing overfitting while maintaining the expressiveness of TDB models as much as possible. It is applicable to most TDB models, while also ensuring tractable computation. Existing regularization methods for KGC rely on minimizing the norms of embeddings to regularize the model [Yang et al., 2014, Lacroix et al., 2018], leading to suboptimal performance. To achieve superior performance, we present an intermediate variables regularization (IVR) approach that minimizes the norms of intermediate variables involved in the processes of computing the predicted tensor of TDB models. Additionally, our approach fully considers the computing ways because different ways of computing the predicted tensor may generate different intermediate variables.

To support the efficacy of our regularization approach, we further provide a theoretical analysis. We prove that our regularization is an upper bound of the overlapped trace norm [Tomioka et al., 2011]. The overlapped trace norm is the sum of the trace norms of the unfolding matrices along each mode of a tensor [Kolda and Bader, 2009], which can be considered as a surrogate measure of the rank of a tensor. Thus, the overlapped trace norm reflects the correlation among entities and relations, which can pose a constraint to jointly entities and relations embeddings learning. In specific, entities and relations in KGs are usually highly correlated. For example, some relations are mutual inverse relations, or a relation may be a composition of another two relations [Zhang et al., 2021]. Through minimizing the upper bound of the overlapped trace norm, we encourage a high correlation among entities and relations, which brings strong regularization and alleviates the overfitting problem.

The main contributions of this paper are listed below:

1. We present a detailed overview of a wide range of TDB models and establish a general form to serve as a foundation for further TDB model analysis.
2. We introduce a new regularization approach for TDB models based on the general form to mitigate the overfitting issue, which is notable for its generality and effectiveness.
3. We provide a theoretical proof of the efficacy of our regularization and validate its practical utility through experiments.

2 Related Work

Tensor Decomposition Based Models Research in KGC has been vast, with TDB models garnering attention due to their superior performance. The two primary TDB approaches that have been extensively studied are CP decomposition [Hitchcock, 1927] and Tucker decomposition [Tucker, 1966]. CP decomposition represents a tensor as a sum of n rank-one tensors, while Tucker decomposition decomposes a tensor into a core tensor and a set of matrices. Kolda and Bader [2009] shows more details about CP decomposition and Tucker decomposition.

Several techniques have been developed for applying CP decomposition and Tucker decomposition in KGC. For instance, Lacroix et al. [2018] employed the original CP decomposition, whereas DistMult [Yang et al., 2014], a variant of CP decomposition, made the embedding matrices of head entities and tail entities identical to simplify the model. SimpleE [Kazemi and Poole, 2018] tackled the problem of independence among the embeddings of head and tail entities within CP decomposition. ComplEx [Trouillon et al., 2017] extended DistMult to the complex space to handle asymmetric relations. QuatE [Zhang et al., 2019] explored hypercomplex space to further enhance KGC. Other techniques

include HoLE [Nickel et al., 2016] proposed by Nickel et al. [2016], which operates on circular correlation, and ANALOGY [Liu et al., 2017], which explicitly utilizes the analogical structures of KGs. Notably, Liu et al. [2017] confirmed that HoLE is equivalent to ComplEx. Additionally, Balažević et al. [2019] introduced TuckER, which is based on the Tucker decomposition and has achieved state-of-the-art performance across various benchmark datasets for KGC. Nickel et al. [2011] proposed a three-way decomposition RESCAL over each relational slice of the tensor. You can refer to [Zhang et al., 2021] or [Ji et al., 2021] for more detailed discussion about KGC models or TDB models.

Regularization Although TDB models are highly expressive [Trouillon et al., 2017, Kazemi and Poole, 2018, Balažević et al., 2019], they can suffer severely from overfitting in practice. Consequently, several regularization approaches have been proposed. A common regularization approach is to apply the squared Frobenius norm to the model parameters [Nickel et al., 2011, Yang et al., 2014, Trouillon et al., 2017]. However, this approach does not correspond to a proper tensor norm, as shown by Lacroix et al. [2018]. Therefore, they proposed a novel regularization method, N3, based on the tensor nuclear 3-norm, which is an upper bound of the tensor nuclear norm. Likewise, Zhang et al. [2020] introduced DURA, a regularization method that exploits the duality of TDB models and distance-based models, and serves as an upper bound of the tensor nuclear 2-norm. However, both N3 and DURA are derived from the CP decomposition, and thus are only applicable to CP and ComplEx models.

3 Methods

In Section 3.1, we begin by providing an overview of existing TDB models and derive a general form for them. Thereafter, we present our intermediate variables regularization (IVR) approach in Section 3.2. Finally, in Section 3.3, we provide theoretical analysis to support the efficacy of our proposed regularization technique.

3.1 General Form

To facilitate the subsequent theoretical analysis, we initially provide a summary of existing TDB models and derive a general form for them. Given a set of entities \mathcal{E} and a set of relations \mathcal{R} , a KG contains a set of triplets $\mathcal{S} = \{(i, j, k)\} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. Let $\mathbf{X} \in \{0, 1\}^{|\mathcal{E}| \times |\mathcal{R}| \times |\mathcal{E}|}$ represent the KG tensor, with $X_{ijk} = 1$ iff $(i, j, k) \in \mathcal{S}$, where $|\mathcal{E}|$ and $|\mathcal{R}|$ denote the number of entities and relations, respectively. Let $\mathbf{H} \in \mathbb{R}^{|\mathcal{E}| \times D}$, $\mathbf{R} \in \mathbb{R}^{|\mathcal{R}| \times D}$ and $\mathbf{T} \in \mathbb{R}^{|\mathcal{E}| \times D}$ be the embedding matrices of head entities, relations and tail entities, respectively, where D is the embedding dimension.

Various TDB models can be attained by partitioning the embedding matrices into P parts. We reshape $\mathbf{H} \in \mathbb{R}^{|\mathcal{E}| \times D}$, $\mathbf{R} \in \mathbb{R}^{|\mathcal{R}| \times D}$ and $\mathbf{T} \in \mathbb{R}^{|\mathcal{E}| \times D}$ into $\mathbf{H} \in \mathbb{R}^{|\mathcal{E}| \times (D/P) \times P}$, $\mathbf{R} \in \mathbb{R}^{|\mathcal{R}| \times (D/P) \times P}$ and $\mathbf{T} \in \mathbb{R}^{|\mathcal{E}| \times (D/P) \times P}$, respectively, where P is the number of parts we partition. For different P , we can get different TDB models.

CP/DistMult Let $P = 1$, CP [Lacroix et al., 2018] can be represented as

$$X_{ijk} = \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:1}, \mathbf{T}_{k:1} \rangle := \sum_{d=1}^{D/P} \mathbf{H}_{id1} \mathbf{R}_{jd1} \mathbf{T}_{kd1}$$

where $\langle \cdot, \cdot, \cdot \rangle$ is the dot product of three vectors. DistMult [Yang et al., 2014], a particular case of CP, which shares the embedding matrices of head entities and tail entities, i.e., $\mathbf{H} = \mathbf{T}$.

ComplEx/HoLE Let $P = 2$, ComplEx [Trouillon et al., 2017] can be represented as

$$X_{ijk} = \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:1}, \mathbf{T}_{k:1} \rangle + \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:1}, \mathbf{T}_{k:2} \rangle + \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:2}, \mathbf{T}_{k:2} \rangle - \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:2}, \mathbf{T}_{k:1} \rangle$$

Liu et al. [2017] proved that HoLE [Nickel et al., 2011] is equivalent to ComplEx.

Simple Let $P = 2$, Simple [Kazemi and Poole, 2018] can be represented as

$$X_{ijk} = \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:1}, \mathbf{T}_{k:2} \rangle + \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:2}, \mathbf{T}_{k:1} \rangle$$

ANALOGY Let $P = 4$, ANALOGY [Liu et al., 2017] can be represented as

$$\begin{aligned} X_{ijk} = & \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:1}, \mathbf{T}_{k:1} \rangle + \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:2}, \mathbf{T}_{k:2} \rangle + \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:3}, \mathbf{T}_{k:3} \rangle + \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:4}, \mathbf{T}_{k:4} \rangle \\ & + \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:3}, \mathbf{T}_{k:4} \rangle - \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:4}, \mathbf{T}_{k:3} \rangle \end{aligned}$$

QuatE Let $P = 4$, QuatE [Zhang et al., 2019] can be represented as

$$\begin{aligned} X_{ijk} = & \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:1}, \mathbf{T}_{k:1} \rangle - \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:2}, \mathbf{T}_{k:1} \rangle - \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:3}, \mathbf{T}_{k:1} \rangle - \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:4}, \mathbf{T}_{k:1} \rangle \\ & + \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:2}, \mathbf{T}_{k:2} \rangle + \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:1}, \mathbf{T}_{k:2} \rangle + \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:4}, \mathbf{T}_{k:2} \rangle - \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:3}, \mathbf{T}_{k:2} \rangle \\ & + \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:3}, \mathbf{T}_{k:3} \rangle - \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:4}, \mathbf{T}_{k:3} \rangle + \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:1}, \mathbf{T}_{k:3} \rangle + \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:2}, \mathbf{T}_{k:3} \rangle \\ & + \langle \mathbf{H}_{i:1}, \mathbf{R}_{j:4}, \mathbf{T}_{k:4} \rangle + \langle \mathbf{H}_{i:2}, \mathbf{R}_{j:3}, \mathbf{T}_{k:4} \rangle - \langle \mathbf{H}_{i:3}, \mathbf{R}_{j:2}, \mathbf{T}_{k:4} \rangle + \langle \mathbf{H}_{i:4}, \mathbf{R}_{j:1}, \mathbf{T}_{k:4} \rangle \end{aligned}$$

Tucker Let $P = D$, Tucker [Balažević et al., 2019] can be represented as

$$X_{ijk} = \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \mathbf{H}_{i1l} \mathbf{R}_{j1m} \mathbf{T}_{k1n}$$

where $\mathbf{W} \in \mathbb{R}^{P \times P \times P}$ is the core tensor.

General Form Through our analysis, we observe that all TDB models can be expressed as a linear combination of several dot product. The key distinguishing factors among these models are the choice of the number of parts P and the core tensor \mathbf{W} . The number of parts P determines the dimensions of the dot products of the embeddings, while the core tensor \mathbf{W} determines the strength of the dot products. It is important to note that Tucker uses a parameter tensor as its core tensor, whereas the core tensors of other models are predetermined constant tensors. Therefore, we can derive a general form of these models as

$$\begin{aligned} X_{ijk} &= \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \langle \mathbf{H}_{i:l}, \mathbf{R}_{j:m}, \mathbf{T}_{k:n} \rangle = \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \left(\sum_{d=1}^{D/P} \mathbf{H}_{idl} \mathbf{R}_{jdm} \mathbf{T}_{kdn} \right) \\ &= \sum_{d=1}^{D/P} \left(\sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \mathbf{H}_{idl} \mathbf{R}_{jdm} \mathbf{T}_{kdn} \right) \end{aligned} \quad (1)$$

or

$$\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d} \quad (2)$$

where \times_n is the mode- n product [Kolda and Bader, 2009], and $\mathbf{W} \in \mathbb{R}^{P \times P \times P}$ is the core tensor, which can be a parameter tensor or a predetermined constant tensor. The general form Eq.(2) can also be considered as a sum of D/P Tucker decompositions, which is also called block-term decomposition [De Lathauwer, 2008]. Eq.(2) is a block-term decomposition with a shared core tensor \mathbf{W} . This general form is easy to understand, facilitates better understanding of TDB models and paves the way for further exploration of TDB models. The general form presents a unified view of TDB models and helps the researchers understand the relationship between different TDB models. Moreover, the general form motivates the researchers to propose new methods and establish unified theoretical frameworks that are applicable to most TDB models. Our proposed regularization in Section 3.2 and the theoretical analysis Section 3.3 are examples of such contributions.

The Number of Parameters and Computational Complexity The parameters of Eq.(2) come from two parts, the core tensor \mathbf{W} and the embedding matrices \mathbf{H} , \mathbf{R} and \mathbf{T} . The number of parameters of the core tensor \mathbf{W} is equal to P^3 if \mathbf{W} is a parameter tensor and otherwise equal to 0. The number of parameters of the embedding matrices is equal to $|\mathcal{E}|D + |\mathcal{R}|D$ if $\mathbf{H} = \mathbf{T}$ and otherwise equal to $2|\mathcal{E}|D + |\mathcal{R}|D$. The computational complexity of Eq.(2) is equal to $\mathcal{O}(DP^2|\mathcal{E}|^2|\mathcal{R}|)$. The larger the number of parts P , the more expressive the model and the more the computation. Therefore, the choice of P is a trade-off between expressiveness and computation.

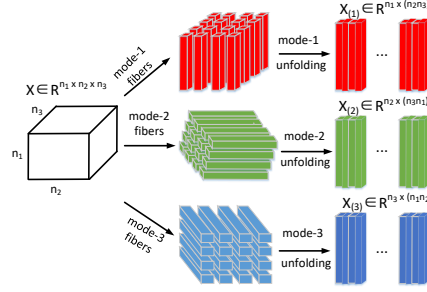


Figure 1: Left shows a 3rd order tensor. Middle describes the corresponding mode- i fibers of the tensor. Fibers are the higher-order analogue of matrix rows and columns. A fiber is defined by fixing every index but one. Right describes the corresponding mode- i unfolding of the tensor. The mode- i unfolding of a tensor arranges the mode- i fibers to be the columns of the resulting matrix.

Tucker and Eq.(2) Tucker [Balažević et al., 2019] also demonstrated that TDB models can be represented as a Tucker decomposition by setting specific core tensors \mathbf{W} . Nevertheless, we must stress that Tucker does not explicitly consider the number of parts P and the core tensor \mathbf{W} , which are pertinent to the number of parameters and computational complexity of TDB models. Moreover, in Appendix A, we demonstrate that the conditions for a TDB model to learn logical rules are also dependent on P and \mathbf{W} . By selecting appropriate P and \mathbf{W} , TDB models can be able to learn symmetry rules, antisymmetry rules, and inverse rules.

3.2 Intermediate Variables Regularization

TDB models are theoretically fully expressive [Trouillon et al., 2017, Kazemi and Poole, 2018, Balažević et al., 2019], which can represent any real-valued tensor. However, TDB models suffer from the overfitting problem in practice [Lacroix et al., 2018]. Therefore, several regularization methods have been proposed, such as squared Frobenius norm method [Yang et al., 2014] and nuclear 3-norm method [Lacroix et al., 2018], which minimize the norms of the embeddings $\{\mathbf{H}, \mathbf{R}, \mathbf{T}\}$ to regularize the model. Nonetheless, merely minimizing the embeddings tends to have suboptimal impacts on the model performance. To enhance the model performance, we introduce a new regularization method that minimizes the norms of the intermediate variables involved in the processes of computing \mathbf{X} . To ensure the broad applicability of our method, our regularization is rooted in the general form of TDB models Eq.(2).

To compute \mathbf{X} in Eq.(2), we can first compute the intermediate variable $\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}$, and then combine $\mathbf{T}_{:d}$ to compute \mathbf{X} . Thus, in addition to minimizing the norm of $\mathbf{T}_{:d}$, we also need to minimize the norm of $\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}$. Since different ways of computing \mathbf{X} can result in different intermediate variables, we fully consider the computing ways of \mathbf{X} . Eq.(2) can also be written as:

$$\mathbf{X} = \sum_{d=1}^{D/P} (\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}) \times_1 \mathbf{H}_{:d},$$

$$\mathbf{X} = \sum_{d=1}^{D/P} (\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}) \times_2 \mathbf{R}_{:d}.$$

Thus, we also need to minimize the norms of intermediate variables $\{\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}, \mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\}$ and $\{\mathbf{H}_{:d}, \mathbf{R}_{:d}\}$. In summary, we should minimize the (power of Frobenius) norms $\{\|\mathbf{H}_{:d}\|_F^\alpha, \|\mathbf{R}_{:d}\|_F^\alpha, \|\mathbf{T}_{:d}\|_F^\alpha\}$ and $\{\|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha, \|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha, \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha\}$, where α is the power of the norms.

Since computing \mathbf{X} is equivalent to computing $\mathbf{X}_{(1)}$ or $\mathbf{X}_{(2)}$ or $\mathbf{X}_{(3)}$, we can also minimize the norms of intermediate variables involved in the processes of computing $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$, where $\mathbf{X}_{(n)}$ is the mode- n unfolding of a tensor \mathbf{X} [Kolda and Bader, 2009]. See Figure 1 for an example

of the notation $\mathbf{X}_{(n)}$. We can represent $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$ as [Kolda and Bader, 2009]:

$$\begin{aligned}\mathbf{X}_{(1)} &= \sum_{d=1}^{D/P} (\mathbf{W} \times_1 \mathbf{H}_{:d})_{(1)} (\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T, \\ \mathbf{X}_{(2)} &= \sum_{d=1}^{D/P} (\mathbf{W} \times_2 \mathbf{R}_{:d})_{(2)} (\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T, \\ \mathbf{X}_{(3)} &= \sum_{d=1}^{D/P} (\mathbf{W} \times_3 \mathbf{T}_{:d})_{(3)} (\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T\end{aligned}$$

where \otimes is the Kronecker product. Thus, the intermediate variables include $\{\mathbf{W} \times_1 \mathbf{H}_{:d}, \mathbf{W} \times_2 \mathbf{R}_{:d}, \mathbf{W} \times_3 \mathbf{T}_{:d}\}$ and $\{\mathbf{T}_{:d} \otimes \mathbf{R}_{:d}, \mathbf{T}_{:d} \otimes \mathbf{H}_{:d}, \mathbf{R}_{:d} \otimes \mathbf{H}_{:d}\}$. Therefore, we should minimize the (power of Frobenius) norms $\{\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha, \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha, \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha\}$ and $\{\|\mathbf{T}_{:d} \otimes \mathbf{R}_{:d}\|_F^\alpha = \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha, \|\mathbf{T}_{:d} \otimes \mathbf{H}_{:d}\|_F^\alpha = \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha, \|\mathbf{R}_{:d} \otimes \mathbf{H}_{:d}\|_F^\alpha = \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha\}$.

Our Intermediate Variables Regularization (IVR) is defined as a combination of all these norms:

$$\begin{aligned}\text{reg}(\mathbf{X}) &= \sum_{d=1}^{D/P} \lambda_1 (\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\ &\quad + \lambda_2 (\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\ &\quad + \lambda_3 (\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha) \\ &\quad + \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha) \quad (3)\end{aligned}$$

where $\{\lambda_i > 0 | i = 1, 2, 3, 4\}$ are the regularization coefficients.

In conclusion, our proposed regularization term is the sum of the norms of variables involved in the different ways of computing the tensor \mathbf{X} .

We can easily get the weighted version of Eq.(3), in which the regularization term corresponding to the sampled training triplets only [Lacroix et al., 2018, Zhang et al., 2020]. For a training triplet (i, j, k) , the weighted version of Eq.(3) is as follows:

$$\begin{aligned}\text{reg}(\mathbf{X}_{ijk}) &= \sum_{d=1}^{D/P} \lambda_1 (\|\mathbf{H}_{id}\|_F^\alpha + \|\mathbf{R}_{jd}\|_F^\alpha + \|\mathbf{T}_{kd}\|_F^\alpha) \\ &\quad + \lambda_2 (\|\mathbf{T}_{kd}\|_F^\alpha \|\mathbf{R}_{jd}\|_F^\alpha + \|\mathbf{T}_{kd}\|_F^\alpha \|\mathbf{H}_{id}\|_F^\alpha + \|\mathbf{R}_{jd}\|_F^\alpha \|\mathbf{H}_{id}\|_F^\alpha) \\ &\quad + \lambda_3 (\|\mathbf{W} \times_1 \mathbf{H}_{id}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{jd}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{kd}\|_F^\alpha) \\ &\quad + \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{jd} \times_3 \mathbf{T}_{kd}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{kd} \times_1 \mathbf{H}_{id}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{id} \times_2 \mathbf{R}_{jd}\|_F^\alpha) \quad (4)\end{aligned}$$

The computational complexity of Eq.(4) is the same as that of Eq.(1), i.e., $\mathcal{O}(DP^2)$, which ensures that our regularization is computationally tractable.

The hyper-parameters λ_i make IVR scalable. We can easily reduce the number of hyper-parameters by setting some of them zero or equal. The hyper-parameters make us able to achieve a balance between performance and efficiency as shown in Section 4.3. We set $\lambda_1 = \lambda_3$ and $\lambda_2 = \lambda_4$ for all models to reduce the number of hyper-parameters. You can refer to Appendix C for more details about the setting of hyper-parameters.

We use the same loss function, multiclass log-loss function, as in [Lacroix et al., 2018]. For a training triplet (i, j, k) , our loss function is

$$\ell(\mathbf{X}_{ijk}) = -\mathbf{X}_{ijk} + \log\left(\sum_{k'=1}^{|\mathcal{E}|} \exp(\mathbf{X}_{ijk'})\right) + \text{reg}(\mathbf{X}_{ijk})$$

At test time, we use $\mathbf{X}_{i,j,:}$ to rank tail entities for a query $(i, j, ?)$.

3.3 Theoretical Analysis

To support the effectiveness of our regularization IVR, we provide a deeper theoretical analysis of its properties. The establishment of the theoretical framework of IVR is inspired by Lemma 1 in

Appendix B, which relates the Frobenius norm and the trace norm of a matrix. Lemma 1 shows that the trace norm of a matrix is an upper bound of a function of several Frobenius norms of intermediate variables, which prompts us to establish the relationship between IVR and trace norm. Based on Lemma 1, we prove that IVR serves as an upper bound for the overlapped trace norm of the predicted tensor, which promotes the low nuclear norm of the predicted tensor to regularize the model.

The overlapped trace norm [Kolda and Bader, 2009] for a 3rd-order tensor is defined as:

$$L(\mathbf{X}; \alpha) := \|\mathbf{X}_{(1)}\|_*^{\alpha/2} + \|\mathbf{X}_{(2)}\|_*^{\alpha/2} + \|\mathbf{X}_{(3)}\|_*^{\alpha/2}$$

where α is the power coefficient in Eq.(3). $\|\mathbf{X}_{(1)}\|_*$, $\|\mathbf{X}_{(2)}\|_*$ and $\|\mathbf{X}_{(3)}\|_*$ are the matrix trace norms of $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$, respectively, which are the sums of singular values of the respective matrices. The matrix trace norm is widely used as a convex surrogate for matrix rank due to the non-differentiability of matrix rank [Goldfarb and Qin, 2014, Lu et al., 2016, Mu et al., 2014]. Thus, $L(\mathbf{X}; \alpha)$ serves as a surrogate for $\text{rank}(\mathbf{X}_{(1)})^{\alpha/2} + \text{rank}(\mathbf{X}_{(2)})^{\alpha/2} + \text{rank}(\mathbf{X}_{(3)})^{\alpha/2}$, where $\text{rank}(\mathbf{X}_{(1)})$, $\text{rank}(\mathbf{X}_{(2)})$ and $\text{rank}(\mathbf{X}_{(3)})$ are the matrix ranks of $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$, respectively. In KGs, each head entity, each relation and each tail entity uniquely corresponds to a row of $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$, respectively. Therefore, $\text{rank}(\mathbf{X}_{(1)})$, $\text{rank}(\mathbf{X}_{(2)})$ and $\text{rank}(\mathbf{X}_{(3)})$ measure the correlation among the head entities, relations and tail entities, respectively. Entities or relations in KGs are highly correlated. For instance, some relations are mutual inverse relations or one relation may be a composition of another two relations [Zhang et al., 2021]. Thus, the overlapped trace norm $L(\mathbf{X}; \alpha)$ can pose a constraint to the embeddings of entities and relations. Minimizing $L(\mathbf{X}; \alpha)$ encourage a high correlation among entities and relations, which brings strong regularization and reduces overfitting. We next establish the relationship between our regularization term Eq.(3) and $L(\mathbf{X}; \alpha)$ by Proposition 1 and Proposition 2. We will prove that Eq.(3) is an upper bound of $L(\mathbf{X}; \alpha)$.

Proposition 1. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, we have

$$\begin{aligned} 2\sqrt{\lambda_1 \lambda_4} L(\mathbf{X}; \alpha) &\leq \sum_{d=1}^{D/P} \lambda_1 (\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\ &\quad + \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha) \end{aligned} \quad (5)$$

If $\mathbf{X}_{(1)} = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^T$, $\mathbf{X}_{(2)} = \mathbf{U}_2 \Sigma_2 \mathbf{V}_2^T$, $\mathbf{X}_{(3)} = \mathbf{U}_3 \Sigma_3 \mathbf{V}_3^T$ are compact singular value decompositions of $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$, $\mathbf{X}_{(3)}$ [Bai et al., 2000], respectively, then there exists a decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, such that the two sides of Eq.(5) equal.

Proposition 2. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, we have

$$\begin{aligned} 2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}) &\leq \sum_{d=1}^{D/P} \lambda_2 (\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\ &\quad + \lambda_3 (\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha) \end{aligned} \quad (6)$$

And there exists some \mathbf{X}' , and for any decomposition of \mathbf{X}' , such that the two sides of Eq.(6) can not achieve equality.

Please refer to Appendix B for the proofs. Proposition 1 establishes that the r.h.s. of Eq.(5) provides a tight upper bound for $2\sqrt{\lambda_1 \lambda_4} L(\mathbf{X}; \alpha)$, while Proposition 2 demonstrates that the r.h.s. of Eq.(4) is an upper bound of $2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}; \alpha)$, but this bound is not always tight. Our proposed regularization term, Eq.(3), combines these two upper limits by adding the r.h.s. of Eq.(5) and the r.h.s. of Eq.(6). As a result, minimizing Eq.(3) can effectively minimize $L(\mathbf{X}; \alpha)$ to regularize the model.

The two sides of Eq.(5) can achieve equality if $P = D$, meaning that the TDB model is Tucker model [Balažević et al., 2019]. Although $L(\mathbf{X}; \alpha)$ may not always serve as a tight lower bound of the r.h.s. of Eq.(5) for TDB models other than Tucker model, it remains a common lower bound for all TDB models. To obtain a more tight lower bound, the exact values of P and W are required. For example, in the case of CP model [Lacroix et al., 2018] ($P = 1$ and $W = 1$), the nuclear 2-norm $\|\mathbf{X}\|_*$ is a more tight lower bound. The nuclear 2-norm is defined as follows:

$$\|\mathbf{X}\|_* := \min \left\{ \sum_{d=1}^D \|\mathbf{H}_{:d1}\|_F \|\mathbf{R}_{:d1}\|_F \|\mathbf{T}_{:d1}\|_F \mid \mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1} \right\}$$

Table 1: Knowledge graph completion results on WN18RR, FB15k-237 and YGAO3-10 datasets.

	WN18RR			FB15k-237			YAGO3-10		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
CP	0.438	0.416	0.485	0.332	0.244	0.507	0.567	0.495	0.696
CP-F2	0.449	0.420	0.506	0.331	0.243	0.507	0.570	0.499	0.699
CP-N3	0.469	0.432	0.541	0.355	0.261	0.542	0.575	0.504	0.703
CP-DURA	0.471	0.433	0.545	0.364	0.269	0.554	0.577	0.504	0.704
CP-IVR	0.478	0.437	0.554	0.365	0.270	0.555	0.579	0.507	0.708
ComplEx	0.464	0.431	0.526	0.348	0.256	0.531	0.574	0.501	0.704
ComplEx-F2	0.467	0.431	0.538	0.349	0.260	0.529	0.576	0.502	0.709
ComplEx-N3	0.491	0.445	0.578	0.367	0.272	0.559	0.577	0.504	0.707
ComplEx-DURA	0.484	0.440	0.571	0.372	0.277	0.563	0.585	0.512	0.714
ComplEx-IVR	0.494	0.449	0.581	0.370	0.275	0.561	0.586	0.515	0.714
Simple	0.443	0.421	0.488	0.337	0.248	0.514	0.565	0.491	0.696
Simple-F2	0.451	0.422	0.506	0.338	0.249	0.514	0.566	0.494	0.699
Simple-IVR	0.470	0.436	0.537	0.357	0.264	0.544	0.578	0.504	0.707
ANALOGY	0.458	0.424	0.526	0.348	0.255	0.530	0.573	0.502	0.703
ANALOGY-F2	0.467	0.434	0.533	0.349	0.258	0.529	0.573	0.501	0.705
ANALOGY-IVR	0.482	0.439	0.568	0.367	0.272	0.558	0.582	0.509	0.713
QuatE	0.460	0.430	0.518	0.349	0.258	0.530	0.566	0.489	0.702
QuatE-F2	0.468	0.435	0.534	0.349	0.259	0.529	0.566	0.489	0.705
QuatE-IVR	0.493	0.447	0.580	0.369	0.274	0.561	0.582	0.509	0.712
TuckER	0.446	0.423	0.490	0.321	0.233	0.498	0.551	0.476	0.689
TuckER-F2	0.449	0.423	0.496	0.327	0.239	0.503	0.566	0.492	0.700
TuckER-IVR	0.501	0.460	0.579	0.368	0.274	0.555	0.581	0.508	0.712

where $\mathbf{W} = 1$. The following proposition establishes the relationship between $\|\mathbf{X}\|_*$ and $L(\mathbf{X}; 2)$:

Proposition 3. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}$, and $\mathbf{W} = 1$, we have

$$2\sqrt{\lambda_1 \lambda_4} L(\mathbf{X}; 2) \leq 6\sqrt{\lambda_1 \lambda_4} \|\mathbf{X}\|_* \leq \sum_{d=1}^D \lambda_1 (\|\mathbf{H}_{:d1}\|_F^2 + \|\mathbf{R}_{:d1}\|_F^2 + \|\mathbf{T}_{:d1}\|_F^{2a}) \\ + \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}\|_F^2 + \|\mathbf{W} \times_3 \mathbf{T}_{:d1} \times_1 \mathbf{H}_{:d1}\|_F^2 + \|\mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1}\|_F^2)$$

Although the r.h.s. of Eq.(6) is not always a tight upper bound for $2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}; \alpha)$ like the r.h.s. of Eq.(5), we observe that minimizing the combination of these two bounds, Eq.(3), can lead to better performance. The reason behind this is that the r.h.s. of Eq.(5) is neither an upper bound nor a lower bound of the r.h.s. of Eq.(6) for all \mathbf{X} . We present Proposition 4 in Appendix B to prove this claim.

4 Experiments

We first introduce the experimental settings in Section 4.1 and show the results in Section 4.2. We next conduct ablation studies in Section 4.3. Finally, we verify the reliability of our proposed upper bounds in Section 4.4. Please refer to Appendix C for more experimental details.

4.1 Experimental Settings

Datasets We evaluate the models on three KGC datasets, WN18RR [Dettmers et al., 2018], FB15k-237 [Toutanova et al., 2015] and YAGO3-10 [Dettmers et al., 2018].

Models We use CP, ComplEx, Simple, ANALOGY, QuatE and TuckER as baselines. We denote CP with squared Frobenius norm method [Yang et al., 2014] as CP-F2, CP with N3 method [Lacroix

Table 2: The results on WN18RR and FB15k-237 datasets with different upper bounds.

	WN18RR			FB15k-237			YAGO3-10		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
TuckER	0.446	0.423	0.490	0.321	0.233	0.498	0.551	0.476	0.689
TuckER-IVR-1	0.497	0.455	0.578	0.366	0.272	0.553	0.576	0.501	0.709
TuckER-IVR-2	0.459	0.423	0.518	0.336	0.241	0.518	0.568	0.493	0.700
TuckER-IVR	0.501	0.460	0.579	0.368	0.274	0.555	0.581	0.508	0.712

Table 3: The results on Kinship dataset with different upper bounds.

	$\ \mathbf{X}_{(1)}\ _*$	$\ \mathbf{X}_{(2)}\ _*$	$\ \mathbf{X}_{(3)}\ _*$	$L(\mathbf{X})$
TuckER	10,719	8,713	11,354	30,786
TuckER-IVR-1	5,711	5,021	6,271	17,003
TuckER-IVR-2	6,145	5,441	6,744	18,330
TuckER-IVR	3,538	3,511	3,988	11,037

et al., 2018] as CP-N3, CP with DURA method [Zhang et al., 2020] as CP-DURA and CP with IVR method as CP-IVR. The notations for other models are similar to the notations for CP.

Evaluation Metrics We use the filtered MRR and Hits@N (H@N) [Bordes et al., 2013] as evaluation metrics and choose the hyper-parameters with the best filtered MRR on the validation set. We run each model three times with different random seeds and report the mean results.

4.2 Results

See Table 1 for the results. For CP and ComplEx, the models that N3 and DURA are suitable, the results show that N3 enhances the models more than F2, and DURA outperforms both F2 and N3, leading to substantial improvements. IVR achieves better performance than DURA on WN18RR dataset and achieves similar performance to DURA on FB15k-237 and YAGO3-10 dataset. For SimpleE, ANALOGY, QuatE, and TuckER, the improvement offered by F2 is minimal, while IVR significantly boosts model performance. In summary, these results demonstrate the effectiveness and generality of IVR.

4.3 Ablation Studies

We conduct ablation studies to examine the effectiveness of the upper bounds. Our notations for the models are as follows: the model with upper bound Eq.(5) is denoted as IVR-1, model with upper bound Eq.(6) as IVR-2, and model with upper bound Eq.(3) as IVR.

See Table 2 for the results. We use TuckER as the baseline. IVR with only 1 regularization coefficient, IVR-1, achieves comparable results with vanilla IVR, which shows that IVR can still perform well with fewer hyper-parameters. IVR-1 outperforms IVR-2 due to the tightness of Eq.(5).

4.4 Upper Bounds

We verify that minimizing the upper bounds can effectively minimize $L(\mathbf{X}; \alpha)$. $L(\mathbf{X}; \alpha)$ can measure the correlation of \mathbf{X} . Lower values of $L(\mathbf{X}; \alpha)$ encourage higher correlations among entities and relations, and thus bring a strong constraint for regularization. Upon training the models, we compute $L(\mathbf{X}; \alpha)$. As computing $L(\mathbf{X}; \alpha)$ for large KGs is impractical, we conduct experiments on a small KG dataset, Kinship [Kok and Domingos, 2007], which consists of 104 entities and 25 relations. We use TuckER as the baseline and compare it against IVR-1 (Eq.(5)), IVR-2 (Eq.(6)), and IVR (Eq.(3)).

See Table 3 for the results. Our results demonstrate that the upper bounds are effective in minimizing $L(\mathbf{X}; \alpha)$. All three upper bounds can lead to a decrease of $L(\mathbf{X}; \alpha)$, achieving better performance (Table 2) by more effective regularization. The $L(\mathbf{X}; \alpha)$ of IVR-1 is smaller than that of IVR-2

because the upper bound in Eq.(5) is tight. IVR, which combines IVR-1 and IVR-2, produces the most reduction of $L(\mathbf{X}; \alpha)$. This finding suggests that combining the two upper bounds can be more effective. Overall, our experimental results confirm the reliability of our theoretical analysis.

5 Conclusion

In this paper, we undertake an analysis of TDB models in KGC. We first offer a summary of TDB models and derive a general form that facilitates further analysis. TDB models often suffer from overfitting, and thus, we propose a regularization based on our derived general form. It is applicable to most TDB models and improve the model performance. We further propose a theoretical analysis to support our regularization and experimentally validate our theoretical analysis. Our regularization is limited to TDB models, hoping that more regularization will be proposed in other types of models, such as translation-based models and neural networks models. We also intend to explore how to apply our regularization to other fields, such as tensor completion [Song et al., 2019].

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2020AAA0106000), the National Natural Science Foundation of China (U19A2079), and the CCCD Key Lab of Ministry of Culture and Tourism.

References

- Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst. *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM, 2000.
- Ivana Balažević, Carl Allen, and Timothy Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, 2019.
- James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
- Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Reexamining low rank matrix factorization for trace norm regularization. *arXiv preprint arXiv:1706.08934*, 2017.
- Lieven De Lathauwer. Decompositions of a higher-order tensor in block terms—part i: Lemmas for partitioned matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1022–1032, 2008.
- Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- Donald Goldfarb and Zhiwei Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.
- Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

- Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4289–4300, 2018.
- Stanley Kok and Pedro Domingos. Statistical predicate invention. In *Proceedings of the 24th international conference on Machine learning*, pages 433–440, 2007.
- Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.
- Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*, pages 2863–2872. PMLR, 2018.
- Hanxiao Liu, Yuexin Wu, and Yiming Yang. Analogical inference for multi-relational embeddings. In *International conference on machine learning*, pages 2168–2178. PMLR, 2017.
- Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5249–5257, 2016.
- Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81. PMLR, 2014.
- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 809–816, 2011.
- Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–48, 2019.
- Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical performance of convex tensor decomposition. *Advances in neural information processing systems*, 24, 2011.
- Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1499–1509, 2015.
- Théo Trouillon, Christopher R Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research*, 18:1–38, 2017.
- Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv e-prints*, pages arXiv-1412, 2014.
- Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2:14–35, 2021.
- Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2735–2745, 2019.
- Zhanqiu Zhang, Jianyu Cai, and Jie Wang. Duality-induced regularizer for tensor factorization based knowledge graph completion. *Advances in Neural Information Processing Systems*, 33, 2020.

A Logical Rules

KGs often involve some logical rules to capture inductive capacity [Zhang et al., 2021]. Thus, we analyze how to design models such that the models can learn the symmetry, antisymmetry and inverse rules. We have derived a general form for TDB models, Eq.(1). Next, we study how to enable Eq.(1) to learn logical rules. We first define the symmetry rules, antisymmetry rules and inverse rules. We denote $X_{ijk} = f(\mathbf{H}_{i:}, \mathbf{R}_{j:}, \mathbf{T}_{k:})$ as $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ for simplicity. We assume $\mathbf{H} = \mathbf{T}$, which mainly aims to reduce overfitting Yang et al. [2014]. Existing TDB models for KGC except CP [Lacroix et al., 2018] all forces $\mathbf{H} = \mathbf{T}$.

A relation r is symmetric if $\forall h, t, (h, r, t) \in \mathcal{S} \rightarrow (t, r, h) \in \mathcal{S}$. A model is able to learn the symmetry rules if

$$\exists \mathbf{r} \in \mathbb{R}^D \wedge \mathbf{r} \neq 0, \forall \mathbf{h}, \mathbf{t} \in \mathbb{R}^D, f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = f(\mathbf{t}, \mathbf{r}, \mathbf{h})$$

A relation r is antisymmetric if $\forall h, t, (h, r, t) \in \mathcal{S} \rightarrow (t, r, h) \notin \mathcal{S}$. A model is able to learn the antisymmetry rules if

$$\exists \mathbf{r} \in \mathbb{R}^D \wedge \mathbf{r} \neq 0, \forall \mathbf{h}, \mathbf{t} \in \mathbb{R}^D, f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -f(\mathbf{t}, \mathbf{r}, \mathbf{h})$$

A relation r_1 is inverse to a relation r_2 if $\forall h, t, (h, r_1, t) \in \mathcal{S} \rightarrow (t, r_2, h) \in \mathcal{S}$. A model is able to learn the inverse rules if

$$\forall \mathbf{r}_1 \in \mathbb{R}^D, \exists \mathbf{r}_2 \in \mathbb{R}^D, \forall \mathbf{h}, \mathbf{t} \in \mathbb{R}^D, f(\mathbf{h}, \mathbf{r}_1, \mathbf{t}) = f(\mathbf{t}, \mathbf{r}_2, \mathbf{h})$$

We restrict $\mathbf{r} \neq 0$ because $\mathbf{r} = 0$ will result in f equal to an identically zero function. By choosing different P and \mathbf{W} , we can define different TDB models as discussed in Section 3.1. Next, we give a theoretical analysis to establish the relationship between logical rules and TDB models.

Theorem 1. Assume a model can be represented as the form of Eq.(1), then a model is able to learn the symmetry rules iff $\text{rank}(\mathbf{W}_{(2)}^T - \mathbf{S}\mathbf{W}_{(2)}^T) < P$. A model is able to learn the antisymmetry rules iff $\text{rank}(\mathbf{W}_{(2)}^T + \mathbf{S}\mathbf{W}_{(2)}^T) < P$. A model is able to learn the inverse rules iff $\text{rank}(\mathbf{W}_{(2)}^T) = \text{rank}([\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T])$, where $\mathbf{S} \in \mathbb{R}^{P^2 \times P^2}$ is a permutation matrix with $\mathbf{S}_{(i-1)P+j, (j-1)P+i} = 1 (i, j = 1, 2, \dots, P)$ and otherwise 0, $[\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T]$ is the concatenation of matrix $\mathbf{W}_{(2)}^T$ and matrix $\mathbf{S}\mathbf{W}_{(2)}^T$.

Proof. According to the symmetry rules, for any triplet (i, j, k) , we have that $f(i, j, k) = f(k, j, i)$, i.e., $f(\mathbf{H}_{i:}, \mathbf{R}_{j:}, \mathbf{T}_{k:}) = f(\mathbf{H}_{k:}, \mathbf{R}_{j:}, \mathbf{T}_{i:}) = f(\mathbf{T}_{k:}, \mathbf{R}_{j:}, \mathbf{H}_{i:})$ (we use $\mathbf{H} = \mathbf{T}$ here). We replace $\mathbf{H}_{i:}$ by \mathbf{h} , replace $\mathbf{R}_{j:}$ by \mathbf{r} , replace $\mathbf{T}_{k:}$ by \mathbf{t} , then we have $f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = f(\mathbf{t}, \mathbf{r}, \mathbf{h})$.

Then we have that

$$\begin{aligned} & \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \langle \mathbf{h}_{:l}, \mathbf{r}_{:m}, \mathbf{t}_{:n} \rangle - \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \langle \mathbf{t}_{:l}, \mathbf{r}_{:m}, \mathbf{h}_{:n} \rangle \\ &= \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{lmn} \langle \mathbf{h}_{:l}, \mathbf{r}_{:m}, \mathbf{t}_{:n} \rangle - \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P \mathbf{W}_{nml} \langle \mathbf{h}_{:l}, \mathbf{r}_{:m}, \mathbf{t}_{:n} \rangle \\ &= \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P (\mathbf{W}_{lmn} - \mathbf{W}_{nml}) \langle \mathbf{h}_{:l}, \mathbf{r}_{:m}, \mathbf{t}_{:n} \rangle \\ &= \sum_{l=1}^P \sum_{m=1}^P \sum_{n=1}^P (\mathbf{W}_{lmn} - \mathbf{W}_{nml}) \mathbf{r}_{:m}^T (\mathbf{h}_{:l} * \mathbf{t}_{:n}) \\ &= \sum_{l=1}^P \sum_{n=1}^P \left(\sum_{m=1}^P (\mathbf{W}_{lmn} - \mathbf{W}_{nml}) \mathbf{r}_{:m}^T \right) (\mathbf{h}_{:l} * \mathbf{t}_{:n}) = 0 \end{aligned}$$

where $*$ is the Hadamard product. Since the above equation holds for any $\mathbf{h}, \mathbf{t} \in \mathbb{R}^D$, we can get

$$\sum_{m=1}^P (\mathbf{W}_{lmn} - \mathbf{W}_{nml}) \mathbf{r}_{:m}^T = \mathbf{0} (l, n = 1, 2, \dots, P)$$

Therefore, a model is able to learn the symmetry rules iff

$$\exists \mathbf{r} \in \mathbb{R}^D \wedge \mathbf{r} \neq \mathbf{0}, \sum_{m=1}^P (\mathbf{W}_{lmn} - \mathbf{W}_{nml}) \mathbf{r}_{:m}^T = \mathbf{0}$$

Therefore, the symmetry rule is transformed into a system of linear equations. This system of linear equations have non-zero solution iff $\text{rank}(\mathbf{W}_{(2)}^T - \mathbf{S}\mathbf{W}_{(2)}^T) < P$. Thus, a model is able to learn the symmetry rules iff $\text{rank}(\mathbf{W}_{(2)}^T - \mathbf{S}\mathbf{W}_{(2)}^T) < P$.

Similarly, a model is able to learn the anti-symmetry rules iff $\text{rank}(\mathbf{W}_{(2)}^T + \mathbf{S}\mathbf{W}_{(2)}^T) < P$.

For the inverse rule:

$$\forall \mathbf{r}_1 \in \mathbb{R}^D, \exists \mathbf{r}_2 \in \mathbb{R}^D, \forall \mathbf{h}, \mathbf{t} \in \mathbb{R}^D, f(\mathbf{h}, \mathbf{r}_1, \mathbf{t}) = f(\mathbf{t}, \mathbf{r}_2, \mathbf{h})$$

a model is able to learn the symmetry rules iff the following equation

$$\mathbf{W}_{(2)}^T \mathbf{r}_1 = \mathbf{S}\mathbf{W}_{(2)}^T \mathbf{r}_2$$

for any $\mathbf{r}_1 \in \mathbb{R}^D$, there exists $\mathbf{r}_2 \in \mathbb{R}^D$ such that the equation holds. Thus, the column vectors of $\mathbf{W}_{(2)}^T$ can be expressed linearly by the column vectors of $\mathbf{S}\mathbf{W}_{(2)}^T$. Since \mathbf{S} is a permutation matrix and $\mathbf{S} = \mathbf{S}^T$, we have that $\mathbf{S}^2 = \mathbf{I}$, thus

$$\mathbf{S}\mathbf{W}_{(2)}^T \mathbf{r}_1 = \mathbf{S}^2 \mathbf{W}_{(2)}^T \mathbf{r}_2 = \mathbf{W}_{(2)}^T \mathbf{r}_2$$

For any $\mathbf{r}_1 \in \mathbb{R}^D$, there exists $\mathbf{r}_2 \in \mathbb{R}^D$ such that the above equation holds. Thus, the column vectors of $\mathbf{S}\mathbf{W}_{(2)}^T$ can be expressed linearly by the column vectors of $\mathbf{W}_{(2)}^T$. Therefore, the column space of $\mathbf{W}_{(2)}^T$ is equivalent to the column space of $\mathbf{S}\mathbf{W}_{(2)}^T$, thus we have

$$\text{rank}(\mathbf{W}_{(2)}^T) = \text{rank}(\mathbf{S}\mathbf{W}_{(2)}^T) = \text{rank}([\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T])$$

Meanwhile, if $\text{rank}(\mathbf{W}_{(2)}^T) = \text{rank}([\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T])$, then the columns of $\mathbf{W}_{(2)}^T$ can be expressed linearly by the columns of $\mathbf{S}\mathbf{W}_{(2)}^T$, thus

$$\forall \mathbf{r}_1 \in \mathbb{R}^D, \exists \mathbf{r}_2 \in \mathbb{R}^D, \mathbf{W}_{(2)}^T \mathbf{r}_1 = \mathbf{S}\mathbf{W}_{(2)}^T \mathbf{r}_2$$

Thus, a model is able to learn the inverse rules iff $\text{rank}(\mathbf{W}_{(2)}^T) = \text{rank}([\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T])$. \square

By this theorem, we only need to judge the relationship between P and the matrix rank about $\mathbf{W}_{(2)}$. ComplEx, Simple, ANALOGYY and QuatE design specific core tensors to make the models enable to learn the logical rules. We can easily verify that these models satisfy the conditions in this theorem. For example, for ComplEx, we have that $P = 2$ and

$$\mathbf{W}_{(2)}^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 1 & 0 \end{pmatrix}, \mathbf{S}\mathbf{W}_{(2)}^T = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, [\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T] = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

$$\text{rank}(\mathbf{W}_{(2)}^T - \mathbf{S}\mathbf{W}_{(2)}^T) = \text{rank}\left(\begin{pmatrix} 0 & 0 \\ 0 & 2 \\ 0 & -2 \\ 0 & 0 \end{pmatrix}\right) = 1 < P = 2$$

$$\text{rank}(\mathbf{W}_{(2)}^T + \mathbf{S}\mathbf{W}_{(2)}^T) = \text{rank}\left(\begin{pmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 2 & 0 \end{pmatrix}\right) = 1 < P = 2$$

$$\text{rank}(\mathbf{W}_{(2)}^T) = \text{rank}([\mathbf{W}_{(2)}^T, \mathbf{S}\mathbf{W}_{(2)}^T]) = 2$$

Thus, ComplEx is able to learn the symmetry rules, antisymmetry rules and inverse rules.

B Proofs

To prove the Proposition 1 and Proposition 2, we first prove the following lemma.

Lemma 1.

$$\|\mathbf{Z}\|_*^\alpha = \min_{\mathbf{Z}=\mathbf{U}\mathbf{V}^T} \frac{1}{2}(\lambda\|\mathbf{U}\|_F^{2\alpha} + \frac{1}{\lambda}\|\mathbf{V}\|_F^{2\alpha})$$

where $\alpha > 0$, $\lambda > 0$ and $\{\mathbf{Z}, \mathbf{U}, \mathbf{V}\}$ are real matrices. If $\mathbf{Z} = \hat{\mathbf{U}}\mathbf{\Sigma}\hat{\mathbf{V}}^T$ is a singular value decomposition of \mathbf{Z} , then equality holds for the choice $\mathbf{U} = \lambda^{\frac{-1}{2\alpha}}\hat{\mathbf{U}}\sqrt{\mathbf{\Sigma}}$ and $\mathbf{V} = \lambda^{\frac{1}{2\alpha}}\hat{\mathbf{V}}\sqrt{\mathbf{\Sigma}}$, where $\sqrt{\mathbf{\Sigma}}$ is the element-wise square root of $\mathbf{\Sigma}$.

Proof. The proof of Lemma 1 is based on the proof of Lemma 9 in [Ciliberto et al., 2017]. Let the singular value decomposition of $\mathbf{Z} \in \mathbb{R}^{m \times n}$ be $\mathbf{Z} = \hat{\mathbf{U}}\mathbf{\Sigma}\hat{\mathbf{V}}^T$, where $\hat{\mathbf{U}} \in \mathbb{R}^{m \times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$, $\hat{\mathbf{V}} \in \mathbb{R}^{n \times r}$, $r = \text{rank}(\mathbf{Z})$, $\hat{\mathbf{U}}^T\hat{\mathbf{U}} = \mathbf{I}_{r \times r}$ and $\hat{\mathbf{V}}^T\hat{\mathbf{V}} = \mathbf{I}_{r \times r}$. We choose any \mathbf{U}, \mathbf{V} such that $\mathbf{Z} = \mathbf{U}\mathbf{V}^T$, then we have $\mathbf{\Sigma} = \hat{\mathbf{U}}^T\mathbf{U}\mathbf{V}^T\hat{\mathbf{V}}$. Moreover, since $\hat{\mathbf{U}}, \hat{\mathbf{V}}$ have orthogonal columns, $\|\hat{\mathbf{U}}^T\mathbf{U}\|_F \leq \|\mathbf{U}\|_F$, $\|\hat{\mathbf{V}}^T\mathbf{V}\|_F \leq \|\mathbf{V}\|_F$. Then

$$\begin{aligned} \|\mathbf{Z}\|_*^\alpha &= \text{Tr}(\mathbf{\Sigma})^\alpha = \text{Tr}(\hat{\mathbf{U}}^T\mathbf{U}\mathbf{V}^T\hat{\mathbf{V}})^\alpha \leq \|\hat{\mathbf{U}}^T\mathbf{U}\|_F^\alpha \|\mathbf{V}^T\hat{\mathbf{V}}\|_F^\alpha \\ &\leq (\sqrt{\lambda}\|\mathbf{U}\|_F^\alpha) (\frac{1}{\sqrt{\lambda}}\|\mathbf{V}\|_F^\alpha) \leq \frac{1}{2}(\lambda\|\mathbf{U}\|_F^{2\alpha} + \frac{1}{\lambda}\|\mathbf{V}\|_F^{2\alpha}) \end{aligned}$$

where the first upper bound is Cauchy-Schwarz inequality and the third upper bound is AM-GM inequality.

Let $\mathbf{U} = \lambda^{\frac{-1}{2\alpha}}\hat{\mathbf{U}}\sqrt{\mathbf{\Sigma}}$ and $\mathbf{V} = \lambda^{\frac{1}{2\alpha}}\hat{\mathbf{V}}\sqrt{\mathbf{\Sigma}}$, we have that

$$\frac{1}{2}(\lambda\|\mathbf{U}\|_F^{2\alpha} + \frac{1}{\lambda}\|\mathbf{V}\|_F^{2\alpha}) = \frac{1}{2}(\|\hat{\mathbf{U}}\sqrt{\mathbf{\Sigma}}\|_F^{2\alpha} + \|\hat{\mathbf{V}}\sqrt{\mathbf{\Sigma}}\|_F^{2\alpha}) = \frac{1}{2}(\|\sqrt{\mathbf{\Sigma}}\|_F^{2\alpha} + \|\sqrt{\mathbf{\Sigma}}\|_F^{2\alpha}) = \|\mathbf{Z}\|_*^\alpha$$

In summary,

$$\|\mathbf{Z}\|_*^\alpha = \min_{\mathbf{Z}=\mathbf{U}\mathbf{V}^T} \frac{1}{2}(\lambda\|\mathbf{U}\|_F^{2\alpha} + \frac{1}{\lambda}\|\mathbf{V}\|_F^{2\alpha})$$

□

Proposition 1. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, we have

$$\begin{aligned} 2\sqrt{\lambda_1\lambda_4}L(\mathbf{X}; \alpha) &\leq \sum_{d=1}^{D/P} \lambda_1(\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\ &\quad + \lambda_4(\|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha) \end{aligned} \quad (7)$$

If $\mathbf{X}_{(1)} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$, $\mathbf{X}_{(2)} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$, $\mathbf{X}_{(3)} = \mathbf{U}_3\mathbf{\Sigma}_3\mathbf{V}_3^T$ are compact singular value decompositions of $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \mathbf{X}_{(3)}$, respectively, then there exists a decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, such that the two sides of Eq.(5) equal, where $P = D$, $\mathbf{H}_{:1} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}}\mathbf{U}_1\sqrt{\mathbf{\Sigma}_1}$, $\mathbf{R}_{:1} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}}\mathbf{U}_2\sqrt{\mathbf{\Sigma}_2}$, $\mathbf{T}_{:1} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}}\mathbf{U}_3\sqrt{\mathbf{\Sigma}_3}$, $\mathbf{W} = \sqrt{\lambda_1/\lambda_4}^{\frac{3}{\alpha}}\mathbf{X} \times_1 \sqrt{\mathbf{\Sigma}_1^{-1}}\mathbf{U}_1^T \times_2 \sqrt{\mathbf{\Sigma}_2^{-1}}\mathbf{U}_2^T \times_3 \sqrt{\mathbf{\Sigma}_3^{-1}}\mathbf{U}_3^T$.

Proof. Let the n -rank [Kolda and Bader, 2009] of $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be (r_1, r_2, r_3) , then $\mathbf{U}_1 \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{U}_2 \in \mathbb{R}^{n_2 \times r_2}$, $\mathbf{U}_3 \in \mathbb{R}^{n_3 \times r_3}$, $\mathbf{\Sigma}_1 \in \mathbb{R}^{r_1 \times r_1}$, $\mathbf{\Sigma}_2 \in \mathbb{R}^{r_2 \times r_2}$, $\mathbf{\Sigma}_3 \in \mathbb{R}^{r_3 \times r_3}$, $\mathbf{W} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

If $\mathbf{X} = \mathbf{0}$, the above proposition is obviously true, we define $\mathbf{0}^{-1} := \mathbf{0}$ here.

For any $\mathbf{X} \neq \mathbf{0}$, since $\mathbf{X}_{(1)} = \sum_{d=1}^{D/P} \mathbf{H}_{:d}(\mathbf{W}_{(1)}(\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T)$, $\mathbf{X}_{(2)} = \sum_{d=1}^{D/P} \mathbf{R}_{:d}(\mathbf{W}_{(2)}(\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T)$, $\mathbf{X}_{(3)} = \sum_{d=1}^{D/P} \mathbf{T}_{:d}(\mathbf{W}_{(3)}(\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T)$, by applying Lemma 1 to $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \mathbf{X}_{(3)}$, we

have that

$$\begin{aligned}
& 2L(\mathbf{X}; \alpha) \\
& \leq \sum_{d=1}^{D/P} \|\mathbf{H}_{:d}(\mathbf{W}_{(1)}(\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T)\|_*^{\alpha/2} + \|\mathbf{R}_{:d}(\mathbf{W}_{(2)}(\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T)\|_*^{\alpha/2} + \|\mathbf{T}_{:d}(\mathbf{W}_{(3)}(\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T)\|_*^{\alpha/2} \\
& \leq \sum_{d=1}^{D/P} \lambda(\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\
& + \frac{1}{\lambda}(\|\mathbf{W}_{(1)}(\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T\|_F^\alpha + \|\mathbf{W}_{(2)}(\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T\|_F^\alpha + \|\mathbf{W}_{(3)}(\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T\|_F^\alpha) \\
& = \sum_{d=1}^{D/P} \lambda(\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\
& + \frac{1}{\lambda}(\|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha)
\end{aligned}$$

Let $\lambda = \sqrt{\lambda_1/\lambda_4}$, we have that

$$\begin{aligned}
2\sqrt{\lambda_1\lambda_4}L(\mathbf{X}; \alpha) & \leq \sum_{d=1}^{D/P} \lambda_1(\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\
& + \lambda_4(\|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d}\|_F^\alpha)
\end{aligned}$$

Since $\mathbf{U}_1^T \mathbf{U}_1 = \mathbf{I}_{r_1 \times r_1}$, $\mathbf{U}_2^T \mathbf{U}_2 = \mathbf{I}_{r_2 \times r_2}$, $\mathbf{U}_3^T \mathbf{U}_3 = \mathbf{I}_{r_3 \times r_3}$, thus we have $\mathbf{X}_{(1)} = \mathbf{U}_1 \mathbf{U}_1^T \mathbf{X}_{(1)}$, $\mathbf{X}_{(2)} = \mathbf{U}_2 \mathbf{U}_2^T \mathbf{X}_{(2)}$, $\mathbf{X}_{(3)} = \mathbf{U}_3 \mathbf{U}_3^T \mathbf{X}_{(3)}$, then

$$\begin{aligned}
\mathbf{X} & = \mathbf{X} \times_1 (\mathbf{U}_1 \mathbf{U}_1^T) \times_2 (\mathbf{U}_2 \mathbf{U}_2^T) \times_3 (\mathbf{U}_3 \mathbf{U}_3^T) \\
& = \mathbf{X} \times_1 (\mathbf{U}_1 \sqrt{\Sigma_1} \sqrt{\Sigma_1^{-1}} \mathbf{U}_1^T) \times_2 (\mathbf{U}_2 \sqrt{\Sigma_2} \sqrt{\Sigma_2^{-1}} \mathbf{U}_2^T) \times_3 (\mathbf{U}_3 \sqrt{\Sigma_3} \sqrt{\Sigma_3^{-1}} \mathbf{U}_3^T) \\
& = \mathbf{X} \times_1 \sqrt{\Sigma_1^{-1}} \mathbf{U}_1^T \times_2 \sqrt{\Sigma_2^{-1}} \mathbf{U}_2^T \times_3 \sqrt{\Sigma_3^{-1}} \mathbf{U}_3^T \times_1 \mathbf{U}_1 \sqrt{\Sigma_1} \times_2 \mathbf{U}_2 \sqrt{\Sigma_2} \times_3 \mathbf{U}_3 \sqrt{\Sigma_3}
\end{aligned}$$

Let $P = D$ and $\mathbf{H}_{:1:} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_1 \sqrt{\Sigma_1}$, $\mathbf{R}_{:1:} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_2 \sqrt{\Sigma_2}$, $\mathbf{T}_{:1:} = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_3 \sqrt{\Sigma_3}$, $\mathbf{W} = \sqrt{\lambda_1/\lambda_4}^{\frac{3}{\alpha}} \mathbf{X} \times_1 \sqrt{\Sigma_1^{-1}} \mathbf{U}_1^T \times_2 \sqrt{\Sigma_2^{-1}} \mathbf{U}_2^T \times_3 \sqrt{\Sigma_3^{-1}} \mathbf{U}_3^T$, then $\mathbf{X} = \mathbf{W} \times_1 \mathbf{H}_{:1:} \times_2 \mathbf{R}_{:1:} \times_3 \mathbf{T}_{:1:}$ is a decomposition of \mathbf{X} . Since

$$\begin{aligned}
\mathbf{X}_{(1)} & = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_1 \sqrt{\Sigma_1} \mathbf{W}_{(1)}(\mathbf{T}_{:1:} \otimes \mathbf{R}_{:1:})^T = \mathbf{U}_1 \sqrt{\Sigma_1} \sqrt{\Sigma_1} \mathbf{V}_1^T \\
\mathbf{X}_{(2)} & = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_2 \sqrt{\Sigma_2} \mathbf{W}_{(2)}(\mathbf{T}_{:1:} \otimes \mathbf{H}_{:1:})^T = \mathbf{U}_2 \sqrt{\Sigma_2} \sqrt{\Sigma_2} \mathbf{V}_2^T \\
\mathbf{X}_{(3)} & = \sqrt{\lambda_1/\lambda_4}^{\frac{-1}{\alpha}} \mathbf{U}_3 \sqrt{\Sigma_3} \mathbf{W}_{(3)}(\mathbf{R}_{:1:} \otimes \mathbf{H}_{:1:})^T = \mathbf{U}_3 \sqrt{\Sigma_3} \sqrt{\Sigma_3} \mathbf{V}_3^T
\end{aligned}$$

thus

$$\begin{aligned}
\mathbf{W}_{(1)}(\mathbf{T}_{:1:} \otimes \mathbf{R}_{:1:})^T & = \sqrt{\lambda_1/\lambda_4}^{\frac{1}{\alpha}} \sqrt{\Sigma_1} \mathbf{V}_1^T \\
\mathbf{W}_{(2)}(\mathbf{T}_{:1:} \otimes \mathbf{H}_{:1:})^T & = \sqrt{\lambda_1/\lambda_4}^{\frac{1}{\alpha}} \sqrt{\Sigma_2} \mathbf{V}_2^T \\
\mathbf{W}_{(3)}(\mathbf{R}_{:1:} \otimes \mathbf{H}_{:1:})^T & = \sqrt{\lambda_1/\lambda_4}^{\frac{1}{\alpha}} \sqrt{\Sigma_3} \mathbf{V}_3^T
\end{aligned}$$

If $P = D$, we have that

$$\begin{aligned}
& \sum_{d=1}^{D/P} \lambda_1(\|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha) \\
& + \lambda_4(\|\mathbf{W}_{(1)}(\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T\|_F^\alpha + \|\mathbf{W}_{(2)}(\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T\|_F^\alpha + \|\mathbf{W}_{(3)}(\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T\|_F^\alpha) \\
& = \lambda_1(\|\mathbf{H}_{:1}\|_F^\alpha + \|\mathbf{R}_{:1}\|_F^\alpha + \|\mathbf{T}_{:1}\|_F^\alpha) \\
& + \lambda_4(\|\mathbf{W} \times_2 \mathbf{R}_{:1} \times_3 \mathbf{T}_{:1}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:1} \times_1 \mathbf{H}_{:1}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{:1} \times_2 \mathbf{R}_{:1}\|_F^\alpha) \\
& = \lambda_1(\|\mathbf{H}_{:1}\|_F^\alpha + \|\mathbf{R}_{:1}\|_F^\alpha + \|\mathbf{T}_{:1}\|_F^\alpha) \\
& + \lambda_4(\|\mathbf{W}_{(1)}(\mathbf{T}_{:1} \otimes \mathbf{R}_{:1})^T\|_F^\alpha + \|\mathbf{W}_{(2)}(\mathbf{T}_{:1} \otimes \mathbf{H}_{:1})^T\|_F^\alpha + \|\mathbf{W}_{(3)}(\mathbf{R}_{:1} \otimes \mathbf{H}_{:1})^T\|_F^\alpha) \\
& = \sqrt{\lambda_1 \lambda_4}(\|\mathbf{U}_1 \sqrt{\Sigma_1}\|_F^\alpha + \|\mathbf{U}_2 \sqrt{\Sigma_2}\|_F^\alpha + \|\mathbf{U}_3 \sqrt{\Sigma_3}\|_F^\alpha) \\
& + \sqrt{\lambda_1 \lambda_4}(\|\sqrt{\Sigma_1} \mathbf{V}_1^T\|_F^\alpha + \|\sqrt{\Sigma_2} \mathbf{V}_2^T\|_F^\alpha + \|\sqrt{\Sigma_3} \mathbf{V}_3^T\|_F^\alpha) \\
& = \sqrt{\lambda_1 \lambda_4}(\|\sqrt{\Sigma_1}\|_F^\alpha + \|\sqrt{\Sigma_2}\|_F^\alpha + \|\sqrt{\Sigma_3}\|_F^\alpha) + \sqrt{\lambda_1 \lambda_4}(\|\sqrt{\Sigma_1}\|_F^\alpha + \|\sqrt{\Sigma_2}\|_F^\alpha + \|\sqrt{\Sigma_3}\|_F^\alpha) \\
& = \sqrt{\lambda_1 \lambda_4}(\|\mathbf{X}_{(1)}\|_*^\alpha + \|\mathbf{X}_{(2)}\|_*^\alpha + \|\mathbf{X}_{(3)}\|_*^\alpha + \|\mathbf{X}_{(1)}\|_*^\alpha + \|\mathbf{X}_{(2)}\|_*^\alpha + \|\mathbf{X}_{(3)}\|_*^\alpha) \\
& = 2L(\mathbf{X}; \alpha)
\end{aligned}$$

□

Proposition 2. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}$, we have

$$\begin{aligned}
2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}) & \leq \sum_{d=1}^{D/P} \lambda_2(\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\
& + \lambda_3(\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha)
\end{aligned} \tag{8}$$

And there exists some \mathbf{X}' , and for any decomposition of \mathbf{X}' , such that the two sides of Eq.(6) can not achieve equality.

Proof. Since $\mathbf{X}_{(1)} = \sum_{d=1}^{D/P} (\mathbf{H}_{:d} \mathbf{W}_{(1)}) (\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T$, $\mathbf{X}_{(2)} = \sum_{d=1}^{D/P} (\mathbf{R}_{:d} \mathbf{W}_{(2)}) (\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T$, $\mathbf{X}_{(3)} = \sum_{d=1}^{D/P} (\mathbf{T}_{:d} \mathbf{W}_{(3)}) (\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T$ and for any matrix \mathbf{A}, \mathbf{B} , $\|\mathbf{A} \otimes \mathbf{B}\|_F^\alpha = \|\mathbf{A}\|_F^\alpha \|\mathbf{B}\|_F^\alpha$, by applying Lemma 1 to $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \mathbf{X}_{(3)}$, we have that

$$\begin{aligned}
& 2L(\mathbf{X}; \alpha) \\
& \leq \sum_{d=1}^{D/P} \|(\mathbf{H}_{:d} \mathbf{W}_{(1)}) (\mathbf{T}_{:d} \otimes \mathbf{R}_{:d})^T\|_*^{\alpha/2} + \|(\mathbf{R}_{:d} \mathbf{W}_{(2)}) (\mathbf{T}_{:d} \otimes \mathbf{H}_{:d})^T\|_*^{\alpha/2} + \|(\mathbf{T}_{:d} \mathbf{W}_{(3)}) (\mathbf{R}_{:d} \otimes \mathbf{H}_{:d})^T\|_*^{\alpha/2} \\
& \leq \sum_{d=1}^{D/P} \lambda(\|\mathbf{T}_{:d} \otimes \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d} \otimes \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d} \otimes \mathbf{H}_{:d}\|_F^\alpha) \\
& + \frac{1}{\lambda}(\|\mathbf{H}_{:d} \mathbf{W}_{(1)}\|_F^\alpha + \|\mathbf{R}_{:d} \mathbf{W}_{(2)}\|_F^\alpha + \|\mathbf{T}_{:d} \mathbf{W}_{(3)}\|_F^\alpha) \\
& = \sum_{d=1}^{D/P} \lambda(\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\
& + \lambda(\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha)
\end{aligned}$$

Let $\lambda = \sqrt{\lambda_2/\lambda_3}$, we have that

$$\begin{aligned}
2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}) & \leq \sum_{d=1}^{D/P} \lambda_2(\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\
& + \lambda_3(\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha)
\end{aligned} \tag{9}$$

Let $\mathbf{X}' \in \mathbb{R}^{2 \times 2 \times 2}$, $\mathbf{X}'_{1,1,1} = \mathbf{X}'_{2,2,2} = 1$ and $\mathbf{X}'_{ijk} = 0$ otherwise. Then

$$\mathbf{X}'_{(1)} = \mathbf{X}'_{(2)} = \mathbf{X}'_{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Thus $\text{rank}(\mathbf{X}'_{(1)}) = \text{rank}(\mathbf{X}'_{(2)}) = \text{rank}(\mathbf{X}'_{(3)}) = 2$. In Lemma 1, a necessary condition of the equality holds is that $\text{rank}(\mathbf{U}) = \text{rank}(\mathbf{V}) = \text{rank}(\mathbf{Z})$. Thus, the equality holds only if $P = D$ and

$$\begin{aligned} \text{rank}(\mathbf{X}'_{(1)}) &= \text{rank}(\mathbf{T}_{:1} \otimes \mathbf{R}_{:1}) = 2, \text{rank}(\mathbf{X}'_{(2)}) = \text{rank}(\mathbf{T}_{:1} \otimes \mathbf{H}_{:1}) = 2 \\ \text{rank}(\mathbf{X}'_{(3)}) &= \text{rank}(\mathbf{R}_{:1} \otimes \mathbf{H}_{:1}) = 2 \end{aligned}$$

Since for any matrix \mathbf{A}, \mathbf{B} , $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A}) \text{rank}(\mathbf{B})$, we have that

$$\text{rank}(\mathbf{T}_{:1}) \text{rank}(\mathbf{R}_{:1}) = 2, \text{rank}(\mathbf{T}_{:1}) \text{rank}(\mathbf{H}_{:1}) = 2, \text{rank}(\mathbf{R}_{:1}) \text{rank}(\mathbf{H}_{:1}) = 2$$

The above equations have no non-negative integer solution, thus there is no decomposition of \mathbf{X}' such that

$$\begin{aligned} 2\sqrt{\lambda_2 \lambda_3} L(\mathbf{X}) &= \sum_{d=1}^{D/P} \lambda_2 (\|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{T}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{R}_{:d}\|_F^\alpha \|\mathbf{H}_{:d}\|_F^\alpha) \\ &+ \lambda_3 (\|\mathbf{W} \times_1 \mathbf{H}_{:d}\|_F^\alpha + \|\mathbf{W} \times_2 \mathbf{R}_{:d}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{:d}\|_F^\alpha) \end{aligned}$$

□

Proposition 3. For any \mathbf{X} , and for any decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}$, we have

$$\begin{aligned} 2\sqrt{\lambda_1 \lambda_4} L(\mathbf{X}; 2) &\leq 6\sqrt{\lambda_1 \lambda_4} \|\mathbf{X}\|_* \leq \sum_{d=1}^D \lambda_1 (\|\mathbf{H}_{:d1}\|_F^2 + \|\mathbf{R}_{:d1}\|_F^2 + \|\mathbf{T}_{:d1}\|_F^2) \\ &+ \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}\|_F^2 + \|\mathbf{W} \times_3 \mathbf{T}_{:d1} \times_1 \mathbf{H}_{:d1}\|_F^2 + \|\mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1}\|_F^2) \end{aligned}$$

where $\mathbf{W} = 1$.

Proof. $\forall \mathbf{X}$, let $S_1 = \{(\mathbf{W}, \mathbf{H}, \mathbf{R}, \mathbf{T}) | \mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\}$, $S_2 = \{(\mathbf{W}, \mathbf{H}, \mathbf{R}, \mathbf{T}) | \mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}, \mathbf{W} = 1\}$. We have that

$$\begin{aligned} &2\sqrt{\lambda_1 \lambda_4} \left(\sum_{d=1}^D \|\mathbf{H}_{:d1}\|_F \|\mathbf{R}_{:d1}\|_F \|\mathbf{T}_{:d1}\|_F \right) \\ &\leq 2 \left(\sum_{d=1}^D \lambda_4 \|\mathbf{R}_{:d1}\|_F^2 \|\mathbf{T}_{:d1}\|_F^2 \right)^{1/2} \left(\sum_{d=1}^D \lambda_1 \|\mathbf{H}_{:d1}\|_F^2 \right)^{1/2} \\ &= \sum_{d=1}^D \lambda_1 \|\mathbf{H}_{:d1}\|_F^2 + \lambda_4 \|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^2 \end{aligned}$$

where $\mathbf{W} = 1$. The equality holds if and only if $\lambda_4 \|\mathbf{R}_{:d1}\|_F^2 \|\mathbf{T}_{:d1}\|_F^2 = \lambda_1 \|\mathbf{H}_{:d1}\|_F^2$, i.e., $\sqrt{\lambda_4/\lambda_1} \|\mathbf{R}_{:d1}\|_2 \|\mathbf{T}_{:d1}\|_2 = \|\mathbf{H}_{:d1}\|_2$. For a CP decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}$, we let $\mathbf{H}'_{:d1} = \sqrt{\frac{\|\mathbf{R}_{:d1}\|_2 \|\mathbf{T}_{:d1}\|_2}{\|\mathbf{H}_{:d1}\|_2}} \mathbf{H}_{:i}$, $\mathbf{R}'_{:i} = \sqrt{\frac{\|\mathbf{H}_{:d1}\|_2}{\|\mathbf{R}_{:d1}\|_2 \|\mathbf{T}_{:d1}\|_2}} \mathbf{R}_{:i}$, $\mathbf{T}'_{:i} = \mathbf{T}_{:i}$ if $\|\mathbf{H}_{:d1}\|_2 \neq 0$, $\|\mathbf{R}_{:d1}\|_2 \neq 0$, $\|\mathbf{T}_{:d1}\|_2 \neq 0$ and otherwise $\mathbf{H}'_{:d} = \mathbf{0}$, $\mathbf{R}'_{:d} = \mathbf{0}$, $\mathbf{T}'_{:d} = \mathbf{0}$. Then $\mathbf{X} = \sum_{d=1}^D \mathbf{W} \times_1 \mathbf{H}'_{:d1} \times_2 \mathbf{R}'_{:d1} \times_3 \mathbf{T}'_{:d1}$ is another CP decomposition of \mathbf{X} and $\|\mathbf{R}'_{:d1}\|_2 \|\mathbf{T}'_{:d1}\|_2 = \|\mathbf{H}'_{:d1}\|_2$. Thus

$$2\sqrt{\lambda_1 \lambda_4} \|\mathbf{X}\|_* = \min_{S_2} \sum_{d=1}^D \lambda_1 \|\mathbf{H}_{:d1}\|_F^2 + \lambda_4 \|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^2$$

Similarly, we can get

$$2\sqrt{\lambda_1 \lambda_4} \|\mathbf{X}\|_* = \min_{S_2} \sum_{d=1}^D \lambda_1 \|\mathbf{R}_{:d1}\|_F^2 + \lambda_4 \|\mathbf{W} \times_2 \mathbf{T}_{:d} \times_3 \mathbf{H}_{:d}\|_F^2$$

$$2\sqrt{\lambda_1\lambda_4}\|\mathbf{X}\|_* = \min_{S_2} \sum_{d=1}^D \lambda_1 \|\mathbf{T}_{:d1}\|_F^2 + \lambda_4 \|\mathbf{W} \times_2 \mathbf{H}_{:d} \times_3 \mathbf{R}_{:d}\|_F^2$$

By Proposition 1, we have that

$$2\sqrt{\lambda_1\lambda_4}\|\mathbf{X}_{(1)}\|_* = \sum_{d=1}^D \lambda_1 \|\mathbf{H}_{:d1}\|_F^2 + \lambda_4 \|\mathbf{W} \times_2 \mathbf{R}_{:d} \times_3 \mathbf{T}_{:d}\|_F^2$$

Since S_2 is a subset of S_1 , we have that

$$\|\mathbf{X}_{(1)}\|_* \leq \|\mathbf{X}\|_*$$

Similarly, we can prove that $\|\mathbf{X}_{(2)}\|_* \leq \|\mathbf{X}\|_*$ and $\|\mathbf{X}_{(3)}\|_* \leq \|\mathbf{X}\|_*$, thus

$$\begin{aligned} 2\sqrt{\lambda_1\lambda_4}L(\mathbf{X}; 2) &\leq 6\sqrt{\lambda_1\lambda_4}\|\mathbf{X}\|_* \leq \sum_{d=1}^D \lambda_1 (\|\mathbf{H}_{:d1}\|_F^2 + \|\mathbf{R}_{:d1}\|_F^2 + \|\mathbf{T}_{:d1}\|_F^2) \\ &\quad + \lambda_4 (\|\mathbf{W} \times_2 \mathbf{R}_{:d1} \times_3 \mathbf{T}_{:d1}\|_F^2 + \|\mathbf{W} \times_3 \mathbf{T}_{:d1} \times_1 \mathbf{H}_{:d1}\|_F^2 + \|\mathbf{W} \times_1 \mathbf{H}_{:d1} \times_2 \mathbf{R}_{:d1}\|_F^2) \end{aligned}$$

□

Proposition 4. For any \mathbf{X} , there exists a decomposition of $\mathbf{X} = \sum_{d=1}^{D/P} \hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}_{:d} \times_2 \hat{\mathbf{R}}_{:d} \times_3 \hat{\mathbf{T}}_{:d}$, such that

$$\begin{aligned} &\sum_{d=1}^{D/P} \lambda_1 (\|\hat{\mathbf{H}}_{:d}\|_F^\alpha + \|\hat{\mathbf{R}}_{:d}\|_F^\alpha + \|\hat{\mathbf{T}}_{:d}\|_F^\alpha) \\ &+ \lambda_4 (\|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}_{:d} \times_3 \hat{\mathbf{T}}_{:d}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}_{:d} \times_1 \hat{\mathbf{H}}_{:d}\|_F^\alpha + \|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}_{:d} \times_2 \hat{\mathbf{R}}_{:d}\|_F^\alpha) \\ &< \sqrt{\lambda_1\lambda_4}/\sqrt{\lambda_2\lambda_3} \sum_{d=1}^{D/P} \lambda_2 (\|\hat{\mathbf{T}}_{:d}\|_F^\alpha \|\hat{\mathbf{R}}_{:d}\|_F^\alpha + \|\hat{\mathbf{T}}_{:d}\|_F^\alpha \|\hat{\mathbf{H}}_{:d}\|_F^\alpha + \|\hat{\mathbf{R}}_{:d}\|_F^\alpha \|\hat{\mathbf{H}}_{:d}\|_F^\alpha) \\ &+ \lambda_3 (\|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}_{:d}\|_F^\alpha + \|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}_{:d}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}_{:d}\|_F^\alpha) \end{aligned}$$

Furthermore, for some \mathbf{X} , there exists a decomposition of \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} \tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}}_{:d} \times_2 \tilde{\mathbf{R}}_{:d} \times_3 \tilde{\mathbf{T}}_{:d}$, such that

$$\begin{aligned} &\sum_{d=1}^{D/P} \lambda_1 (\|\tilde{\mathbf{H}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{R}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{T}}_{:d}\|_F^\alpha) \\ &+ \lambda_4 (\|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}}_{:d} \times_3 \tilde{\mathbf{T}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}}_{:d} \times_1 \tilde{\mathbf{H}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}}_{:d} \times_2 \tilde{\mathbf{R}}_{:d}\|_F^\alpha) \\ &> \sqrt{\lambda_1\lambda_4}/\sqrt{\lambda_2\lambda_3} \sum_{d=1}^{D/P} \lambda_2 (\|\tilde{\mathbf{T}}_{:d}\|_F^\alpha \|\tilde{\mathbf{R}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{T}}_{:d}\|_F^\alpha \|\tilde{\mathbf{H}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{R}}_{:d}\|_F^\alpha \|\tilde{\mathbf{H}}_{:d}\|_F^\alpha) \\ &+ \lambda_3 (\|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}}_{:d}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}}_{:d}\|_F^\alpha) \end{aligned}$$

Proof. To simplify the notations, we denote $\hat{\mathbf{H}}_{:d}$ as $\hat{\mathbf{H}}$, $\hat{\mathbf{R}}_{:d}$ as $\hat{\mathbf{R}}$ and $\hat{\mathbf{T}}_{:d}$ as $\hat{\mathbf{T}}$. To prove the inequality, we only need to prove

$$\begin{aligned} &\sqrt{\lambda_1/\lambda_4} (\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) \\ &+ \sqrt{\lambda_4/\lambda_1} (\|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}} \times_3 \hat{\mathbf{T}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}} \times_2 \hat{\mathbf{R}}\|_F^\alpha) \\ &> \sqrt{\lambda_2/\lambda_3} (\|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha) \\ &+ \sqrt{\lambda_3/\lambda_2} (\|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}\|_F^\alpha) \end{aligned}$$

Let $c_1 = \sqrt{\lambda_1\lambda_3}/\sqrt{\lambda_1\lambda_3}$, for any \mathbf{X} , $\mathbf{X} = \sum_{d=1}^{D/P} (c_1^{-3}\hat{\mathbf{W}}) \times_1 (c_1\hat{\mathbf{H}}_{:d}) \times_2 (c_1\hat{\mathbf{R}}_{:d}) \times_3 (c_1\hat{\mathbf{T}}_{:d})$ is a decomposition of \mathbf{X} , thus we only need to prove

$$\begin{aligned} &c_2 (\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2} (\|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}} \times_3 \hat{\mathbf{T}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}} \times_2 \hat{\mathbf{R}}\|_F^\alpha) \\ &> c_2 (\|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2} (\|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}\|_F^\alpha) \end{aligned}$$

where $c_2 = \sqrt{\lambda_1^2 \lambda_3} / \sqrt{\lambda_2 \lambda_4}$.

If $\mathbf{X} \in \mathbb{R}^{1 \times 1 \times 1}$, let $c = \mathbf{X}^2 + 100$, $\hat{\mathbf{H}} = \hat{\mathbf{R}} = \hat{\mathbf{T}} = c$, $\hat{\mathbf{W}} = \frac{\mathbf{X}}{c^3}$, we have

$$\begin{aligned} & c_2(\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}} \times_3 \hat{\mathbf{T}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}} \times_2 \hat{\mathbf{R}}\|_F^\alpha) \\ &= c_2(\mathbf{X}^2 + 100)^2 + \frac{1}{c_2} \frac{\mathbf{X}^2}{(\mathbf{X}^2 + 100)^2} \\ &< c_2(\|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}\|_F^\alpha) \\ &= c_2(\mathbf{X}^2 + 100)^4 + \frac{1}{c_2} \frac{\mathbf{X}^2}{(\mathbf{X}^2 + 100)^4} \end{aligned}$$

For any $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $\max\{n_1, n_2, n_3\} > 1$, let $\hat{\mathbf{W}}$ be a diagonal tensor, i.e., $\hat{W}_{i,j,k} = 1$ if $i = j = k$ and otherwise 0 and let $\hat{\mathbf{W}} \in \mathbb{R}^{n_1 n_2 n_3 \times n_1 n_2 n_3}$, $\hat{\mathbf{H}} \in \mathbb{R}^{n_1 \times n_1 n_2 n_3}$, $\hat{\mathbf{R}} \in \mathbb{R}^{n_2 \times n_1 n_2 n_3}$, $\hat{\mathbf{T}} \in \mathbb{R}^{n_3 \times n_1 n_2 n_3}$, $\hat{H}_{i,m} = \mathbf{X}_{ijk}$, $\hat{R}_{j,m} = 1$, $\hat{T}_{k,m} = 1$, $m = (i-1)n_2 n_3 + (j-1)n_3 + k$ and otherwise 0. An example of $\mathbf{X} \in \mathbb{R}^{2 \times 2 \times 2}$ is as follows:

$$\begin{aligned} \hat{\mathbf{H}} &= \begin{pmatrix} \mathbf{X}_{1,1,1} & \mathbf{X}_{1,1,2} & \mathbf{X}_{1,2,1} & \mathbf{X}_{1,2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{X}_{2,1,1} & \mathbf{X}_{2,1,2} & \mathbf{X}_{2,2,1} & \mathbf{X}_{2,2,2} \end{pmatrix} \\ \hat{\mathbf{R}} &= \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \hat{\mathbf{T}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{aligned}$$

Thus $\mathbf{X} = \hat{\mathbf{W}} \times_1 \hat{\mathbf{H}} \times_2 \hat{\mathbf{R}} \times_3 \hat{\mathbf{T}}$ is a decomposition of \mathbf{X} , for $\mathbf{X} = (c_1^{-3} \hat{\mathbf{W}}) \times_1 (c_1 \hat{\mathbf{H}}) \times_2 (c_1 \hat{\mathbf{R}}) \times_3 (c_1 \hat{\mathbf{T}})$ we have

$$\begin{aligned} & c_2(\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}} \times_3 \hat{\mathbf{T}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}} \times_2 \hat{\mathbf{R}}\|_F^\alpha) \\ &= c_2(\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{W}}_{(1)}(\hat{\mathbf{T}} \otimes \hat{\mathbf{R}})^T\|_F^\alpha + \|\hat{\mathbf{W}}_{(2)}(\hat{\mathbf{T}} \otimes \hat{\mathbf{H}})^T\|_F^\alpha + \|\hat{\mathbf{W}}_{(3)}(\hat{\mathbf{R}} \otimes \hat{\mathbf{H}})^T\|_F^\alpha) \\ &< \frac{1}{c_2}(\|(\hat{\mathbf{T}} \otimes \hat{\mathbf{R}})^T\|_F^\alpha + \|(\hat{\mathbf{T}} \otimes \hat{\mathbf{H}})^T\|_F^\alpha + \|(\hat{\mathbf{R}} \otimes \hat{\mathbf{H}})^T\|_F^\alpha) + c_2(\|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha) \\ &= c_2(\|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{H}} \hat{\mathbf{W}}_{(1)}\|_F^\alpha + \|\hat{\mathbf{R}} \hat{\mathbf{W}}_{(2)}\|_F^\alpha + \|\hat{\mathbf{T}} \hat{\mathbf{W}}_{(3)}\|_F^\alpha) \\ &= c_2(\|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{T}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{R}}\|_F^\alpha \|\hat{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2}(\|\hat{\mathbf{W}} \times_1 \hat{\mathbf{H}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_2 \hat{\mathbf{R}}\|_F^\alpha + \|\hat{\mathbf{W}} \times_3 \hat{\mathbf{T}}\|_F^\alpha) \end{aligned}$$

For any $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, let $\tilde{\mathbf{W}} = \mathbf{X} / 2\sqrt{2}$, $\tilde{\mathbf{H}} = \sqrt{2} \mathbf{I}_{n_1 \times n_1}$, $\tilde{\mathbf{R}} = \sqrt{2} \mathbf{I}_{n_2 \times n_2}$, $\tilde{\mathbf{T}} = \sqrt{2} \mathbf{I}_{n_3 \times n_3}$, thus

$$\begin{aligned} & c_2(\|\tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2}(\|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}} \times_3 \tilde{\mathbf{T}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}} \times_1 \tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}} \times_2 \tilde{\mathbf{R}}\|_F^\alpha) \\ &= c_2(2n_1 + 2n_2 + 2n_3) + \frac{1}{c_2}(\|\mathbf{X}_{(1)}\|_F^2 / 2 + \|\mathbf{X}_{(2)}\|_F^2 / 2 + \|\mathbf{X}_{(3)}\|_F^2 / 2) \\ & c_2(\|\tilde{\mathbf{T}}\|_F^\alpha \|\tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{T}}\|_F^\alpha \|\tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{R}}\|_F^\alpha \|\tilde{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2}(\|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}}\|_F^\alpha) \\ &= c_2(4n_3 n_2 + 4n_3 n_1 + 4n_2 n_1) + \frac{1}{c_2}(\|\mathbf{X}_{(1)}\|_F^2 / 4 + \|\mathbf{X}_{(2)}\|_F^2 / 4 + \|\mathbf{X}_{(3)}\|_F^2 / 4) \end{aligned}$$

Thus if $\|\mathbf{X}_{(1)}\|_F^2 > 16c_2^2(n_3 n_2 + n_3 n_1 + n_2 n_1)$, we have

$$\begin{aligned} & c_2(\|\tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{T}}\|_F^\alpha) + \frac{1}{c_2}(\|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}} \times_3 \tilde{\mathbf{T}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}} \times_1 \tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}} \times_2 \tilde{\mathbf{R}}\|_F^\alpha) \\ &> c_2(\|\tilde{\mathbf{T}}\|_F^\alpha \|\tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{T}}\|_F^\alpha \|\tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{R}}\|_F^\alpha \|\tilde{\mathbf{H}}\|_F^\alpha) + \frac{1}{c_2}(\|\tilde{\mathbf{W}} \times_1 \tilde{\mathbf{H}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_2 \tilde{\mathbf{R}}\|_F^\alpha + \|\tilde{\mathbf{W}} \times_3 \tilde{\mathbf{T}}\|_F^\alpha) \end{aligned}$$

end the proof. \square

C Experimental details

Datasets The statistics of the datasets, WN18RR, FB15k-237, YAGO3-10 and Kinship, are shown in Table 4.

Table 4: The statistics of the datasets.

Dataset	#entity	#relation	#train	#valid	#test
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
YGAO3-10	123,188	37	1,079,040	5,000	5,000
Kinship	104	25	8,548	1,069	1,069

Table 5: The settings for total embedding dimension D and number of parts P .

	WN18RR		FB15k-237		YAGO3-10	
	D	P	D	P	D	P
CP	2000	1	2000	1	1000	1
ComplEx	4000	2	4000	2	2000	2
Simple	4000	2	4000	2	2000	2
ANALOGY	4000	4	4000	4	2000	4
QuatE	4000	4	4000	4	2000	4
Tucker	256	1	256	1	256	1

Table 6: The settings for power α and regularization coefficients λ_i .

	WN18RR			FB15k-237			YAGO3-10		
	α	λ_1	λ_2	α	λ_1	λ_2	α	λ_1	λ_2
CP	3.0	05	0.07	2.25	0.005	0.07	2.25	0.0007	0.005
ComplEx	3.0	0.05	0.03	2.25	0.005	0.07	2.25	0.0005	0.007
Simple	3.0	0.03	0.07	2.25	0.005	0.07	2.25	0.0007	0.007
ANALOGY	3.0	0.01	0.07	2.25	0.005	0.07	2.25	0.0007	0.007
QuatE	3.0	0.07	0.03	2.25	0.005	0.05	2.25	0.0005	0.005
Tucker	2.0	0.01	0.03	2.0	0.001	0.01	2.0	0.0003	0.003

Evaluation Metrics $MR = \frac{1}{N} \sum_{i=1}^N \text{rank}_i$, where rank_i is the rank of i th triplet in the test set and N is the number of the triplets. Lower MR indicates better performance.

$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i}$. Higher MRR indicates better performance.

$\text{Hits@N} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{rank}_i \leq N)$, where $\mathbb{I}(\cdot)$ is the indicator function. Hits@N is the ratio of the ranks that no more than N , Higher Hits@N indicates better performance.

Hyper-parameters We use a heuristic approach to choose the hyper-parameters and reduce the computation cost with the help of Hyperopt, a hyper-parameter optimization framework based on TPE [Bergstra et al., 2011].

For the hyper-parameter α , we search for the best α in $\{2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5\}$ with fixed hyper-parameters $\lambda_i = 0.0001$.

For the hyper-parameter λ_i , we set $\lambda_1 = \lambda_3$ and $\lambda_2 = \lambda_4$ for all models to reduce the number of hyper-parameters because we notice that the first row of Eq.(4) $\|H_{id}\|_F^\alpha + \|R_{jd}\|_F^\alpha + \|T_{kd}\|_F^\alpha$ is equal to the third row of Eq.(4) $\|W \times_1 H_{id}\|_F^\alpha + \|W \times_2 R_{jd}\|_F^\alpha + \|W \times_3 T_{kd}\|_F^\alpha$ and the second row of Eq.(4) $\|T_{kd}\|_F^\alpha \|R_{jd}\|_F^\alpha + \|T_{kd}\|_F^\alpha \|H_{id}\|_F^\alpha + \|R_{jd}\|_F^\alpha \|H_{id}\|_F^\alpha$ is equal to the

Table 7: The results on WN18RR dataset with different α .

α	2.0	2.25	2.5	2.75	3.0	3.25	3.5
MRR	0.483	0.486	0.485	0.486	0.494	0.486	0.487
H@1	0.443	0.445	0.441	0.442	0.449	0.443	0.444
H@10	0.556	0.564	0.573	0.572	0.581	0.572	0.570

fourth row of Eq.(4) $\|\mathbf{W} \times_2 \mathbf{R}_{jd} \times_3 \mathbf{T}_{kd}\|_F^\alpha + \|\mathbf{W} \times_3 \mathbf{T}_{kd} \times_1 \mathbf{H}_{id}\|_F^\alpha + \|\mathbf{W} \times_1 \mathbf{H}_{id} \times_2 \mathbf{R}_{jd}\|_F^\alpha$ for CP and ComplEx. We then search the regularization coefficients λ_1 and λ_2 in $\{0.001, 0.003, 0.005, 0.007, 0.01, 0.03, 0.05, 0.07\}$ for WN18RR dataset and FB15k-237 dataset, and search λ_1 and λ_2 in $\{0.0001, 0.0003, 0.0005, 0.0007, 0.001, 0.003, 0.05, 0.007\}$ for YAGO3-10 dataset. Thus, we need 64 runs for each model to find the best λ_i . To further reduce the number of runs, we use Hyperopt, a hyper-parameter optimization framework based on TPE [Bergstra et al., 2011], to tune hyper-parameters. In our experiments, we only need 20 runs to find the best λ_i .

We use Adagrad [Duchi et al., 2011] with learning rate 0.1 as the optimizer. We set the batch size to 100 for WN18RR dataset and FB15k-237 dataset and 1000 for YAGO3-10 dataset. We train the models for 200 epochs. The settings for total embedding dimension D and number of parts P are shown in Table 5. The settings for power α and regularization coefficients λ_i are shown in Table 6.

Random Initialization We run each model three times with different random seeds and report the mean results. We do not report the error bars because our model has very small errors with respect to random initialization. The standard deviations of the results are very small. For example, the standard deviations of MRR, H@1 and H@10 of CP model with our regularization are 0.00037784, 0.00084755 and 0.00058739 on WN18RR dataset, respectively. This indicates that our model is not sensitive to the random initialization.

The hyper-parameter α We analyze the impact of the hyper-parameter, the power of the Frobenius norm α . We run experiments on WN18RR dataset with ComplEx model. We set α to $\{2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5\}$. See Table 7 for the results.

The results show that the performance generally increases as α increases and then decreases as α increases. The best α for WN18RR dataset is 3.0. Therefore, we should set a more appropriate α value to obtain better performance.

The hyper-parameter λ_i We analyze the impact of the hyper-parameter, the regularization coefficient λ_i . We run experiments on WN18RR dataset with ComplEx model. We set λ_i to $\{0.001, 0.003, 0.005, 0.007, 0.01, 0.03, 0.05, 0.07, 0.1\}$. See Table 8 and Table 9 for the results.

The experimental results show that the model performance first increases and then decreases with the increase of λ_i , without any oscillation. Thus, we can choose suitable regularization coefficients to prevent overfitting while maintaining the expressiveness of TDB models as much as possible.

The Number of Parts P In Section 3.1, we show that the number of parts P can affect the expressiveness and computation. Thus, we study the impact of P on the model performance. We evaluate the model on WN18RR dataset. We set the total embedding dimension D to 256, and set the parts P to $\{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. See Table 10 for the results. The time is the AMD Ryzen 7 4800U CPU running time on the test set.

The results show that the model performance generally improves and the running time generally increases as P increases. Thus, the larger the part P , the more expressive the model and the more the computation.

A Pseudocode for IVR We present a pseudocode of our method in Alg.(1).

Table 8: The performance of ComplEx on WN18RR dataset with different λ_1 .

λ_1	0.001	0.003	0.005	0.007	0.01	0.03	0.05	0.07	0.1
MRR	0.473	0.475	0.476	0.476	0.480	0.487	0.491	0.486	0.467
H@1	0.435	0.436	0.436	0.436	0.439	0.442	0.445	0.436	0.411
H@10	0.545	0.555	0.555	0.557	0.562	0.576	0.583	0.585	0.570

Table 9: The performance of ComplEx on WN18RR dataset with different λ_2 .

λ_2	0.001	0.003	0.005	0.007	0.01	0.03	0.05	0.07	0.1
MRR	0.464	0.464	0.464	0.465	0.465	0.467	0.471	0.473	0.472
H@1	0.429	0.430	0.430	0.430	0.432	0.432	0.435	0.437	0.435
H@10	0.528	0.528	0.529	0.529	0.530	0.536	0.540	0.540	0.540

Table 10: The results on WN18RR dataset with different P .

P	1	2	4	8	16	32	64	128	256
MRR	0.437	0.449	0.455	0.455	0.463	0.466	0.485	0.497	0.501
H@1	0.405	0.413	0.413	0.415	0.425	0.428	0.446	0.456	0.460
H@10	0.499	0.520	0.536	0.533	0.536	0.539	0.565	0.574	0.579
Time	1.971s	2.096s	2.111s	2.145s	2.301s	2.704s	3.520s	6.470s	16.336s

Algorithm 1 A pseudocode for IVR**Input:**Core tensor: \mathbf{W} , embeddings: $(\mathbf{H}, \mathbf{R}, \mathbf{T})$, triplet: (i, j, k) .Regularization coefficients: $\lambda_l (l = 1, 2, 3, 4)$, the number of parts P . power coefficients: α **Output:**Output of model Eq.(1): \mathbf{X}_{ijk} , IVR regularization Eq.(6): $\text{reg}(\mathbf{X}_{ijk})$ 1: Initialization: $\text{reg} := 0, \mathbf{x}_{1d} := \mathbf{H}_{id}, \mathbf{x}_{2d} := \mathbf{R}_{jd}, \mathbf{x}_{3d} := \mathbf{T}_{kd}$:2: $\text{reg} := \text{reg} + \sum_{d=1}^{D/P} \lambda_1 (\|\mathbf{x}_{1d}\|_F^\alpha + \|\mathbf{x}_{2d}\|_F^\alpha + \|\mathbf{x}_{3d}\|_F^\alpha)$ 3: $\text{reg} := \text{reg} + \sum_{d=1}^{D/P} \lambda_2 (\|\mathbf{x}_{1d}\|_F^\alpha \|\mathbf{x}_{2d}\|_F^\alpha + \|\mathbf{x}_{1d}\|_F^\alpha \|\mathbf{x}_{3d}\|_F^\alpha + \|\mathbf{x}_{2d}\|_F^\alpha \|\mathbf{x}_{3d}\|_F^\alpha)$ 4: $\mathbf{x}_{1d} := \mathbf{W} \times_1 \mathbf{H}_{id}, \mathbf{x}_{2d} := \mathbf{W} \times_2 \mathbf{R}_{jd}, \mathbf{x}_{3d} := \mathbf{W} \times_3 \mathbf{T}_{kd}$:5: $\text{reg} := \text{reg} + \sum_{d=1}^{D/P} \lambda_3 (\|\mathbf{x}_{1d}\|_F^\alpha + \|\mathbf{x}_{2d}\|_F^\alpha + \|\mathbf{x}_{3d}\|_F^\alpha)$ 6: $\mathbf{x}_{1d} := \mathbf{x}_{1d} \times_2 \mathbf{R}_{jd}, \mathbf{x}_{2d} := \mathbf{x}_{2d} \times_3 \mathbf{T}_{kd}, \mathbf{x}_{3d} := \mathbf{x}_{3d} \times_1 \mathbf{H}_{id}$:7: $\text{reg} := \text{reg} + \sum_{d=1}^{D/P} \lambda_4 (\|\mathbf{x}_{1d}\|_F^\alpha + \|\mathbf{x}_{2d}\|_F^\alpha + \|\mathbf{x}_{3d}\|_F^\alpha)$ 8: $\mathbf{x} := \sum_{d=1}^{D/P} \mathbf{x}_{1d} \times_3 \mathbf{T}_{kd}$:9: **return:** \mathbf{x}, reg

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our claims reflect well the effect of our proposed regularization on knowledge graph completion.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We shortly discuss the limitations in the Conclusion Section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We have provided the detailed assumptions and proofs in the Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided the experimental details in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes] (See Appendix C Paragraph "Random Initialization")

Justification: We have provided the experimental details in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have provided the information about the statistical significance of the experiments in Appendix C Paragraph "Random Initialization".

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We have provided the related compute resources in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: Our research complies in all respects with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Our paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.