PAC-Bayesian Generalization Bounds for Knowledge Graph Representation Learning

Jaejun Lee [1] **Minsung Hwang** [1] **Joyce Jiyoung Whang** [1]

Abstract

While a number of knowledge graph representation learning (KGRL) methods have been proposed over the past decade, very few theoretical analyses have been conducted on them. In this paper, we present the first PAC-Bayesian generalization bounds for KGRL methods. To analyze a broad class of KGRL models, we propose a generic framework named ReED (Relation-aware Encoder-Decoder), which consists of a relation-aware message passing encoder and a triplet classification decoder. Our ReED framework can express at least 15 different existing KGRL models, including not only graph neural network-based models such as R-GCN and CompGCN but also shallow-architecture models such as RotatE and ANALOGY. Our generalization bounds for the ReED framework provide theoretical grounds for the commonly used tricks in KGRL, e.g., parameter-sharing and weight normalization schemes, and guide desirable design choices for practical KGRL methods. We empirically show that the critical factors in our generalization bounds can explain actual generalization errors on three real-world knowledge graphs.

1. Introduction

Knowledge graphs consist of entities and relations where a known fact is represented as a triplet of a head entity, a relation, and a tail entity, e.g., (`Washington_DC`, `Capital_Of`, `USA`). Since real-world knowledge graphs do not exhaustively represent all known facts, knowledge graph completion aims to predict missing facts in knowledge graphs. That is, given an incomplete knowledge graph, the goal is to add missing triplets by finding plausible combinations of the entities and relations. This task can be cast into a triplet classification problem where a model determines whether a given triplet is plausible or not [55].

Knowledge graph representation learning (KGRL) has been considered one of the most effective approaches for knowledge graph completion [8; 25; 60; 61; 11; 22]. By learning representations of entities and relations in a knowledge graph, KGRL methods compute a score for each triplet and determine whether the given triplet is likely true or false based on the score [56]. While shallow-architecture models such as RotatE [45] and ANALOGY [26] also work well in practice, many recently proposed methods [42; 52; 43; 23] utilize graph neural networks (GNNs) [58] or message passing neural networks (MPNNs) [15] to improve the performance by adding a neural message passing encoder.

While KGRL techniques are widely used in many real-life applications [55], very few theoretical analyses have been conducted on them. Recently, Barcelo et al. [2] and Huang et al. [17] have extended the Weisfeiler-Lehman (WL) test [57] to multi-relational graphs to investigate the expressive power of GNNs for knowledge graphs; the expressive power indicates how well a model can distinguish graphs with different structures [59]. On the other hand, the generalization bound indicates how successfully a model solves a given task on the entire dataset compared to its performance on a training set [14; 34]. Though a number of KGRL methods have been proposed over the past decade, the generalization bounds of KGRL have rarely been studied.

PAC (Probably Approximately Correct) learning theory provides fundamental tools for analyzing the generalization bounds [50]; the generalization bounds have been explored using different PAC-based approaches, such as the VC dimension-based [51], the Rademacher complexity-based [3], and the PAC-Bayesian approaches [31; 30]. There have been some studies about the generalization bounds for deep neural networks [5; 4; 35] or for GNNs on standard graphs [18; 28; 13; 38; 62] but not for knowledge graphs. Regarding the generalization bounds of GNNs, Liao et al. [24] have shown that the PAC-Bayesian approach can make a tighter bound than the other approaches [41; 14].

In this paper, we present the first PAC-Bayesian generalization bounds for KGRL methods. To comprehensively

[1]School of Computing, KAIST, Daejeon, South Korea. Correspondence to: Joyce Jiyoung Whang <jjwhang@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

represent and analyze various KGRL models and their possible variants, we design the **Re**lation-aware **E**ncoder-**D**ecoder (ReED) framework consisting of the **R**elation-**A**ware **M**essage-**P**assing (RAMP) encoder and a triplet classification decoder; ReED is a generic framework representing at least 15 different KGRL methods, including both GNN-based and shallow-architecture models. We derive concrete generalization bounds for ReED by proposing a transductive PAC-Bayesian approach for a deterministic triplet classifier, which extends the analyses of Bégin et al. [6] and Neyshabur et al. [35]. We also empirically show the increasing and decreasing trends of the actual generalization errors regarding the critical factors in the generalization bounds using three real-world knowledge graphs. Our theoretical generalization bounds and the empirical observations provide useful guidelines for designing practical KGRL methods. Our contributions can be summarized as:

- We propose a novel ReED framework representing various KGRL methods. Our RAMP encoder in ReED is a comprehensive neural encoder for KGRL that can express models such as CompGCN [52] and R-GCN [42].
- We formulate two types of triplet classification decoders in ReED to cover a wide range of KGRL methods; ReED can represent TransR [25], RotatE [45], DistMult [60], ANALOGY [26], and so forth.
- We prove the generalization bounds for the ReED framework by unrolling two-step recursions for adequately modeling the interactions between relation and entity representations. Our work is the first study about PAC-Bayesian generalization bounds for KGRL.
- We analyze our theoretical findings from a practical model design perspective and empirically show the relationships between the critical factors in the theoretical bounds and the actual generalization errors.[1]

2. Knowledge Graph Completion via Triplet Classification

Given a knowledge graph, the goal of knowledge graph completion is to add plausible triplets by finding appropriate combinations of the entities and relations. We consider a standard transductive knowledge graph completion [8; 45] where it is assumed that all entities and relations are observed during training, and a model predicts the plausibility of new combinations of the entities and relations. This can be viewed as a triplet classification problem, where a model determines whether a given triplet is true or false [55]. For example, assume a knowledge graph contains entities, `Washington_DC`, `USA`, and `Vienna`, and relations `Capital_Of` and `Contains`. Consider two triplets (`USA`, `Contains`, `Washington_DC`) and (`Vienna`, `Capital_Of`, `USA`) that are missing in the given knowledge graph; the former triplet is true and the latter triplet is false. Like this, for new combinations of entities and relations, a triplet classification method predicts whether a given triplet is true or false and then adds only ones considered to be true for knowledge graph completion.

Let us consider a fully observed knowledge graph where the labels of all triplets are known and represent it as $G = (\mathcal{V}, \boldsymbol{X}_{\text{ent}}, \mathcal{R}, \boldsymbol{X}_{\text{rel}}, \mathcal{E})$ where $\mathcal{V}$ is a set of entities, $\boldsymbol{X}_{\text{ent}} \in \mathbb{R}^{|\mathcal{V}|\times d_0}$ is a matrix of entity features, d_0 is the dimension of an entity feature vector, $\mathcal{R}$ is a set of relations, $\boldsymbol{X}_{\text{rel}} \in \mathbb{R}^{|\mathcal{R}|\times d_0'}$ is a matrix of relation features, d_0' is the dimension of a relation feature vector, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ such that $\mathcal{E} := \mathcal{E}^+ \cup \mathcal{E}^-$ where $\mathcal{E}^+$ is a set of true triplets and $\mathcal{E}^-$ is a set of false triplets. In practice, if the entity and relation features are unavailable, we can use one-hot encoding for $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$. Each triplet $(h, r, t) \in \mathcal{E}$ is associated with its label $y_{hrt} \in \{0, 1\}$ such that $y_{hrt} = 1$ if $(h, r, t) \in \mathcal{E}^+$ and $y_{hrt} = 0$ if $(h, r, t) \in \mathcal{E}^-$. By sampling triplets from $\mathcal{E}$ without replacement, we create a set of training triplets denoted by $\widehat{\mathcal{E}}$. As a result, we create a training knowledge graph $\widehat{G} = (\mathcal{V}, \boldsymbol{X}_{\text{ent}}, \mathcal{R}, \boldsymbol{X}_{\text{rel}}, \widehat{\mathcal{E}})$. Note that $\mathcal{E}$ is the full triplet set and $\widehat{\mathcal{E}}$ is the training triplet set. A triplet classification method is trained with $\widehat{G}$ at training time and predicts y_{hrt} of a triplet $(h, r, t) \in \mathcal{E}$ at inference time.

Notation For an entity v and a matrix $\boldsymbol{M}$, let $\boldsymbol{M}[v, :]$ denote the row of $\boldsymbol{M}$ corresponding to v. Also, for a relation r, $\boldsymbol{M}[r, :]$ indicates the row of $\boldsymbol{M}$ corresponding to r. Given a triplet (h, r, t), $\boldsymbol{M}[t, h]$ indicates the element of $\boldsymbol{M}$ at the row corresponding to t and the column corresponding to h. Let a matrix with a superscript (l) indicate the matrix at the l-th layer and L be the total number of layers in the RAMP encoder. Let d_l and d_l' denote the dimension of entity and relation representations at the l-th layer, respectively. Given a vector $\mathbf{x}$, $\text{diag}(\mathbf{x})$ is a diagonal matrix whose diagonal is $\mathbf{x}$. Also, $\mathbf{0}_{n_1\times n_2}$ is an all-zero matrix of size $n_1 \times n_2$. All vectors are row vectors. More details are in Appendix A.

3. ReED Framework for Knowledge Graph Representation Learning

Our ReED framework consists of the RAMP encoder and a triplet classification decoder, where we introduce two different types of decoders: the translational distance decoder and the semantic matching decoder.

3.1. Relation-Aware Message-Passing (RAMP) Encoder

Many recently proposed KGRL methods employ GNNs [52] or MPNNs [42] to learn representations of entities by aggregating representations of the neighboring entities and relations. We formulate our RAMP encoder in Definition 3.1 to

[1] Our code and data are available at `https://github.com/bdi-lab/ReED`

Table 1: Our RAMP encoder can represent R-GCN [42], WGCN [43], and CompGCN [52] by appropriately setting the functions and matrices in Definition 3.1.

	ϕ	ρ, ψ	$\boldsymbol{W}_r^{(l)}$	$\boldsymbol{U}_r^{(l)}$	$\boldsymbol{S}_r^{(l)}[v,:]$
R-GCN [42]	ReLU	identity	$\boldsymbol{W}_r^{(l)}$	$\mathbf{0}_{d'_{l-1}\times d_l}$	$\frac{1}{c_{v,r}}\boldsymbol{A}_r[v,:]$
WGCN [43]	Tanh	identity	$\alpha_r^{(l)}\boldsymbol{W}_0^{(l)}$	$\mathbf{0}_{d'_{l-1}\times d_l}$	$\boldsymbol{A}_r[v,:]$
CompGCN (Sub) [52]	Tanh	identity	$\boldsymbol{W}_{\lambda(r)}^{(l)}$	$-\boldsymbol{W}_{\lambda(r)}^{(l)}$	$\boldsymbol{A}_r[v,:]$
CompGCN (Mult) [52]	Tanh	identity	$\text{diag}\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\boldsymbol{W}_{\lambda(r)}^{(l)}$	$\mathbf{0}_{d'_{l-1}\times d_l}$	$\boldsymbol{A}_r[v,:]$

represent the aggregation process in a general form that can subsume many existing KGRL encoders.

Definition 3.1 (RAMP Encoder for KGRL)**.** Given a training knowledge graph $\widehat{G} = (\mathcal{V}, \boldsymbol{X}_{\text{ent}}, \mathcal{R}, \boldsymbol{X}_{\text{rel}}, \widehat{\mathcal{E}})$, for $l \in \{1, 2, \ldots, L\}$, the RAMP encoder is defined by

$$\boldsymbol{H}^{(0)} = \boldsymbol{X}_{\text{ent}}, \boldsymbol{R}^{(0)} = \boldsymbol{X}_{\text{rel}}$$
$$\boldsymbol{M}_r^{(l)}[v,:] = \left[\boldsymbol{H}^{(l-1)}[v,:] \quad \boldsymbol{R}^{(l-1)}[r,:]\right], \qquad v \in \mathcal{V}, r \in \mathcal{R}$$
$$\boldsymbol{H}^{(l)} = \phi\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)} + \rho\left(\sum_{r\in\mathcal{R}} \boldsymbol{S}_r^{(l)}\psi(\boldsymbol{M}_r^{(l)})\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\right)$$
$$\boldsymbol{R}^{(l)} = \boldsymbol{R}^{(l-1)}\boldsymbol{U}_0^{(l)}$$

where $\boldsymbol{H}^{(l)} \in \mathbb{R}^{|\mathcal{V}|\times d_l}$ is a matrix of entity representations, $\boldsymbol{R}^{(l)} \in \mathbb{R}^{|\mathcal{R}|\times d'_l}$ is a matrix of relation representations, $\boldsymbol{M}_r^{(l)} \in \mathbb{R}^{|\mathcal{V}|\times(d_{l-1}+d'_{l-1})}$ is a matrix for concatenating the entity and relation representations for each relation r, $\boldsymbol{W}_0^{(l)} \in \mathbb{R}^{d_{l-1}\times d_l}$, $\boldsymbol{W}_r^{(l)} \in \mathbb{R}^{d_{l-1}\times d_l}$, $\boldsymbol{U}_r^{(l)} \in \mathbb{R}^{d'_{l-1}\times d_l}$, and $\boldsymbol{U}_0^{(l)} \in \mathbb{R}^{d'_{l-1}\times d'_l}$ are learnable projection matrices, ϕ, ρ, ψ are Lipschitz-continuous activation functions with Lipschitz constants $C_\phi, C_\rho, C_\psi \geq 0$ and $\phi(\mathbf{0}) = \mathbf{0}, \rho(\mathbf{0}) = \mathbf{0}, \psi(\mathbf{0}) = \mathbf{0}$ where $\mathbf{0}$ is a zero vector, and $\boldsymbol{S}_r^{(l)} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is a graph diffusion matrix of $\widehat{G}$ for relation $r \in \mathcal{R}$.

In Definition 3.1, an entity's representation is updated based on the entity and relation representations of its neighbors which are defined per relation using a relation-specific graph diffusion matrix $\boldsymbol{S}_r^{(l)}$ for $r \in \mathcal{R}$. The graph diffusion matrices are constructed by decoupling the training triplets based on the relations so that $\boldsymbol{S}_r^{(l)}$ represents the connections between entities with relation r. A simple way to define $\boldsymbol{S}_r^{(l)}$ is to consider an adjacency matrix $\boldsymbol{A}_r \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ for each relation r such that $\boldsymbol{A}_r[t,h] = 1$ if $(h,r,t) \in \widehat{\mathcal{E}}^+$, $\boldsymbol{A}_r[t,h] = 0$ otherwise. When $\boldsymbol{S}_r^{(l)}[v,:] := \boldsymbol{A}_r[v,:]$, it implies that a model uses a sum aggregator in aggregating neighbors' representations. On the other hand, $\boldsymbol{S}_r^{(l)}$ can also be set to a degree-normalized adjacency matrix, i.e., $\boldsymbol{S}_r^{(l)}[v,:] := \boldsymbol{A}_r[v,:]/\,\texttt{deg}(v)$ where $\texttt{deg}(v)$ is the degree of v; this implies a model uses a mean aggregator.

Several well-known GNN-based KGRL encoders can be considered as special cases of the RAMP encoder. For example, Table 1 shows that our RAMP encoder can represent R-GCN [42], WGCN [43], and CompGCN [52] with two different composition operators, subtraction (Sub) and multiplication (Mult). The key is to appropriately set the activation functions ϕ, ρ, ψ, the projection matrices $\boldsymbol{W}_r^{(l)}$ and $\boldsymbol{U}_r^{(l)}$, and the graph diffusion matrices $\boldsymbol{S}_r^{(l)}$. In Table 1, $c_{v,r}$ for R-GCN is a problem-specific normalization constant defined in [42] and $c_{v,r}$ is usually set to be the number of neighbors of v connected by r. Also, for WGCN, $\alpha_r^{(l)}$ is a relation-specific parameter [43]. For CompGCN, $\lambda(r)$ is a function that categorizes the relation r into one of the normal, inverse, and self connections defined in [52]. Also, in CompGCN, the dimensions of the entity and relation representations should be the same for all layers. The RAMP encoder can also represent CompGCN with a circular correlation operator, which is omitted for brevity in Table 1. More details are described in Appendix B.1.

3.2. Triplet Classification Decoder

Using the entity and relation representations returned by our RAMP encoder (i.e., $\boldsymbol{H}^{(L)}$ and $\boldsymbol{R}^{(L)}$), we design a triplet classification decoder to compute the scores of each triplet for determining whether a given triplet is true or false. While $\boldsymbol{H}^{(L)}$ and $\boldsymbol{R}^{(L)}$ are assumed to come from our RAMP encoder in general, i.e., $L > 0$, we can also skip the RAMP encoder and directly apply our triplet classification decoder. When the RAMP encoder is bypassed, i.e., $L = 0$, we assume $\boldsymbol{H}^{(0)} := \boldsymbol{X}_{\text{ent}}, \boldsymbol{R}^{(0)} := \boldsymbol{X}_{\text{rel}}$.

Let $f_{\mathbf{w}} : \mathcal{V}\times\mathcal{R}\times\mathcal{V} \to \mathbb{R}^2$ denote a triplet classifier with the parameters $\mathbf{w}$. Given a triplet (h,r,t), the classifier assigns two different scores, each of which is stored in $f_{\mathbf{w}}(h,r,t)[0]$ and $f_{\mathbf{w}}(h,r,t)[1]$, where the former is proportional to the likelihood of (h,r,t) being false, and the latter is proportional to the likelihood of (h,r,t) being true.

Depending on how the interactions between entity and relation representations are modeled for computing the scores, we design two decoders: the translational distance (TD) decoder and the semantic matching (SM) decoder. The terms, 'translational distance' and 'semantic matching',

Table 2: Translational Distance Decoder in Definition 3.2 can represent TransR [25] and RotatE [45] and Semantic Matching Decoder in Definition 3.3 can represent DistMult [60] and ANALOGY [26] by appropriately setting the projection matrices.

Decoder	Model	Projection Matrices Setup		
Translational Distance	TransR [25]	$\overline{\boldsymbol{W}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\boldsymbol{F}_r^{\langle j\rangle}$	$\boldsymbol{V}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\boldsymbol{F}_r^{\langle j\rangle}$	$\overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{rel}}^{\langle j\rangle}$
	RotatE [45]	$\overline{\boldsymbol{W}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\begin{bmatrix}\boldsymbol{P}_r^{\langle j\rangle} & \boldsymbol{Q}_r^{\langle j\rangle}\\ -\boldsymbol{Q}_r^{\langle j\rangle} & \boldsymbol{P}_r^{\langle j\rangle}\end{bmatrix}$	$\boldsymbol{V}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}$	$\overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{0}_{d'_L\times d_{L+1}}$
Semantic Matching	DistMult [60]	$\overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\left(\text{diag}\left(\boldsymbol{R}^{(L)}[r,:]\boldsymbol{T}_{\text{rel}}^{\langle j\rangle}\right)\right)\boldsymbol{T}_{\text{ent}}^{\langle j\rangle\top}$		
	ANALOGY [26]	$\overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\boldsymbol{B}_r^{\langle j\rangle}\boldsymbol{T}_{\text{ent}}^{\langle j\rangle\top}$ where	$\boldsymbol{B}_{r_1}^{\langle j\rangle}\boldsymbol{B}_{r_2}^{\langle j\rangle} = \boldsymbol{B}_{r_2}^{\langle j\rangle}\boldsymbol{B}_{r_1}^{\langle j\rangle}, \forall r_1, r_2 \in \mathcal{R}$	

have been also used in [55]. In the TD decoder, the scores of (h, r, t) are computed by the distance between h and t after a relation-specific translation is carried out. On the other hand, in the SM decoder, the score of (h, r, t) is computed by the similarity between the individual components of the triplet. Definition 3.2 defines the TD decoder.

Definition 3.2 (Translational Distance Decoder)**.** For a triplet (h, r, t), the TD decoder computes $f_{\mathbf{w}}(h, r, t)[j]$ for $j \in \{0, 1\}$ using the following formulation:

$$f_{\mathbf{w}}(h,r,t)[j] = \\ -\|\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{W}}_r^{\langle j\rangle} + \boldsymbol{R}^{(L)}[r,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\boldsymbol{V}_r^{\langle j\rangle}\|_2$$

where $\overline{\boldsymbol{W}}_r^{\langle j\rangle}, \boldsymbol{V}_r^{\langle j\rangle} \in \mathbb{R}^{d_L\times d_{L+1}}$ and $\overline{\boldsymbol{U}}_r^{\langle j\rangle} \in \mathbb{R}^{d'_L\times d_{L+1}}$ are learnable projection matrices and d_{L+1} is the dimension of the final entity and relation representations.

Note that $\overline{\boldsymbol{W}}_r^{\langle j\rangle}$ and $\boldsymbol{V}_r^{\langle j\rangle}$ carry out the relation-specific translation for the head and tail entity, respectively, whereas $\overline{\boldsymbol{U}}_r^{\langle j\rangle}$ is a projection matrix for relations. When $L = 0$, $\boldsymbol{H}^{(0)}$ and $\boldsymbol{R}^{(0)}$ are fixed to non-learnable matrices, $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$, respectively. By appropriately setting $\overline{\boldsymbol{W}}_r^{\langle j\rangle}$, $\boldsymbol{V}_r^{\langle j\rangle}$, and $\overline{\boldsymbol{U}}_r^{\langle j\rangle}$ in Definition 3.2, we can represent existing knowledge graph embedding methods as special cases of our TD decoder. Table 2 shows how we should set the three projection matrices in Definition 3.2 to present TransR [25] and RotatE [45]. To allow our decoder to express existing simple knowledge graph embedding methods having no encoder, we introduce two learnable matrices $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle} \in \mathbb{R}^{d_L\times\bar{d}}$ and $\boldsymbol{T}_{\text{rel}}^{\langle j\rangle} \in \mathbb{R}^{d'_L\times\bar{d}'}$ which are only used for specializing our decoder to a particular existing model. When we do not need to simulate an existing model, we can simply drop $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle}$ and $\boldsymbol{T}_{\text{rel}}^{\langle j\rangle}$. For TransR, the entity projection matrices are set to $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\boldsymbol{F}_r^{\langle j\rangle}$ where $\boldsymbol{F}_r^{\langle j\rangle} \in \mathbb{R}^{\bar{d}\times d_{L+1}}$, and we set $\bar{d}' = d_{L+1}$. Also, the relation projection matrix is not defined per relation but shared across all relations. In RotatE, the embedding vectors are originally defined in a complex space [45]. To represent RotatE in our framework, we separately handle the real part and the imaginary part of an embedding vector and concatenate them to represent the whole embedding vector. Let us define $\boldsymbol{R}^{(L)}[r,:]\boldsymbol{T}_{\text{rel}}^{\langle j\rangle} := \begin{bmatrix}\mathbf{p}_r^{\langle j\rangle} & \mathbf{q}_r^{\langle j\rangle}\end{bmatrix}$ where $\mathbf{p}_r^{\langle j\rangle}$ indicates the real part and $\mathbf{q}_r^{\langle j\rangle}$ indicates the imaginary part. Note that $(\mathbf{p}_r^{\langle j\rangle}[i])^2 + (\mathbf{q}_r^{\langle j\rangle}[i])^2 = 1$ for $i \in \{0, 1, \cdots, \frac{\bar{d}'}{2}-1\}$. Also, we define $\boldsymbol{P}_r^{\langle j\rangle} := \text{diag}\left(\mathbf{p}_r^{\langle j\rangle}\right)$ and $\boldsymbol{Q}_r^{\langle j\rangle} := \text{diag}\left(\mathbf{q}_r^{\langle j\rangle}\right)$, and set $\bar{d} = \bar{d}' = d_{L+1}$ to represent RotatE as a special case of our TD decoder. Definition 3.2 can also represent TransE [8], TransH [56], and PairRE [10], which are described in Appendix B.2.

We also define the SM decoder in Definition 3.3.

Definition 3.3 (Semantic Matching Decoder)**.** For a triplet (h, r, t), the SM decoder computes $f_{\mathbf{w}}(h, r, t)[j]$ for $j \in \{0, 1\}$ using the following formulation:

$$f_{\mathbf{w}}(h,r,t)[j] = \boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle}(\boldsymbol{H}^{(L)}[t,:])^\top$$

where $\overline{\boldsymbol{U}}_r^{\langle j\rangle} \in \mathbb{R}^{d_L\times d_L}$ is a relation-specific learnable projection matrix.

When $L > 0$, the entity representation matrix $\boldsymbol{H}^{(L)}$ is the output of our RAMP encoder and is multiplied to the relation-specific learnable projection matrix to score a triplet. Table 2 shows that DistMult [60] and ANALOGY [26] are special cases of Definition 3.3. For DistMult, we set $\bar{d} = \bar{d}'$ in $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle}$ and $\boldsymbol{T}_{\text{rel}}^{\langle j\rangle}$. For ANALOGY, we define $\boldsymbol{B}_r^{\langle j\rangle} \in \mathbb{R}^{\bar{d}\times\bar{d}}$ which is a normal matrix. Definition 3.3 can also represent RESCAL [36], HolE [37], ComplEx [49], SimplE [19], and QuatE [61], which are described in Appendix B.3.

3.3. Expressing Existing KGRL Methods Using ReED

ReED can represent many existing KGRL methods ranging from simple shallow-architecture models [26; 45] to neural encoder-based models [42; 52]. In ReED, a triplet classification decoder can be either combined with the RAMP encoder (i.e., $L > 0$) or used standalone (i.e., $L = 0$). Also,

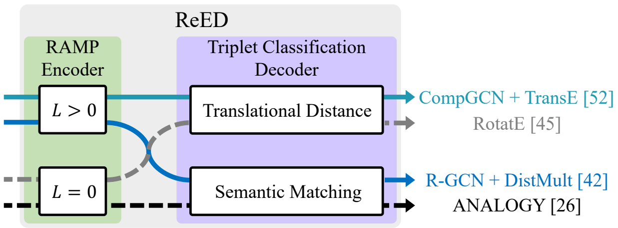

Figure 1: Using different instantiations and combinations of the RAMP encoder and the triplet classification decoder, ReED can express many existing KGRL methods.

the triplet classification decoder can be either the TD or SM decoder. Using different combinations of the RAMP encoder and the decoder, ReED can express various KGRL models, as illustrated in Figure 1. While only four examples are presented in Figure 1, ReED can express more diverse KGRL methods using different instantiations and configurations of the RAMP encoder (Section 3.1) and the triplet classification decoder (Section 3.2).

4. Generalization Bounds for ReED

By proposing a PAC-Bayesian approach for deterministic triplet classifiers for knowledge graph completion, we present PAC-Bayesian generalization bounds of ReED.

4.1. Transductive PAC-Bayesian Approach for Deterministic Triplet Classifiers

The PAC-Bayesian generalization bound relies on the KL divergence of a posterior distribution $\mathcal{Q}$ on a hypothesis space from a prior distribution $\mathcal{P}$ independent of the training set, where $\mathcal{P}$ indicates prior knowledge about a given problem and $\mathcal{Q}$ is learned by a learning algorithm [31; 30]. While the PAC-Bayesian approach was originally designed for analyzing stochastic models [30], our triplet classifiers are deterministic models, i.e., the model parameters are fixed after training. Let us consider a deterministic triplet classifier $f_{\mathbf{w}}(h,r,t)$ which assigns scores of the labels 0 and 1 for $(h,r,t) \in \mathcal{Z}$, where $\mathcal{Z}$ is a finite set of triplets. To gauge the risk of the classifier, a γ-margin loss is defined in Definition 4.1, where the loss is taken into account when the score of the ground-truth label y_{hrt} of a triplet (h,r,t) is less than or equal to that of the other label with the margin of γ. Note that the margin loss is one of the most commonly used loss functions in KGRL [55].

Definition 4.1 (γ-Margin Loss of Triplet Classifier)**.** Given a finite triplet set $\mathcal{Z} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$, for any $\gamma > 0$ and a triplet classifier $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \rightarrow \mathbb{R}^2$ with parameters $\mathbf{w}$ that assigns scores for the label 0 and 1 for $(h,r,t) \in \mathcal{Z}$, a γ-margin loss is defined as

$$\mathcal{L}_{\gamma,\mathcal{Z}}(f_{\mathbf{w}}) = \frac{1}{|\mathcal{Z}|}\sum_{(h,r,t)\in\mathcal{Z}} \mathbb{1}[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \leq \gamma + f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]]$$

where $f_{\mathbf{w}}(h,r,t)[0]$ is the score for label 0, $f_{\mathbf{w}}(h,r,t)[1]$ is the score for label 1, y_{hrt} is the ground-truth label of (h,r,t), and $\mathbb{1}[\cdot]$ is an indicator function.

When $\mathcal{Z}$ is set to the training triplet set $\widehat{\mathcal{E}}$, then $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ is referred to as an empirical loss. On the other hand, a classification loss of a triplet classifier is defined as Definition 4.2.

Definition 4.2 (Classification Loss of Triplet Classifier)**.** Given a finite triplet set $\mathcal{Z} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$, for a triplet classifier $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \rightarrow \mathbb{R}^2$ with parameters $\mathbf{w}$ that assigns scores for the label 0 and 1 for $(h,r,t) \in \mathcal{Z}$, the classification loss is defined as

$$\mathcal{L}_{0,\mathcal{Z}}(f_{\mathbf{w}}) = \frac{1}{|\mathcal{Z}|}\sum_{(h,r,t)\in\mathcal{Z}} \mathbb{1}[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \leq f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]]$$

When $\mathcal{Z}$ is set to the full triplet set $\mathcal{E}$, then $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ is referred to as an expected loss. The generalization bound is defined as the upper bound of the difference between the expected loss $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ and the empirical loss $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$. The generalization bound hints at the level of discrepancy in the model performance between the full and training sets.

As described in Section 2, the full triplet set $\mathcal{E}$ is finite, and the training triplets in $\widehat{\mathcal{E}}$ are sampled from $\mathcal{E}$ without replacement; thus, we need a transductive PAC-Bayesian analysis, which assumes the full set is finite. We derive Theorem 4.3 from Corollary 7 in Bégin et al. [6] by extending it to a deterministic triplet classifier. A main idea of the proof of Theorem 4.3 is to add random perturbations [35] to the fixed parameters $\mathbf{w}$. The proof is available in Appendix C.

Theorem 4.3 (Transductive PAC-Bayesian Generalization Bound for a Deterministic Triplet Classifier)**.** *Let $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \rightarrow \mathbb{R}^2$ be a deterministic triplet classifier with parameters $\mathbf{w}$, and $\mathcal{P}$ be any prior distribution on $\mathbf{w}$. Let us consider the finite full triplet set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$. We construct a posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ by adding any random perturbation $\ddot{\mathbf{w}}$ to $\mathbf{w}$ such that $\mathbb{P}(\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$. Then, for any $\gamma, \delta > 0$, with probability $1-\delta$ over the choice of a training triplet set $\widehat{\mathcal{E}}$ drawn from the full triplet set $\mathcal{E}$ (such that $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$ and $|\mathcal{E}| \geq 40$) without replacement, for any $\mathbf{w}$, we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}$$

where $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ is defined in Definition 4.1, $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ is defined in Definition 4.2, $D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P})$ is the

KL-divergence of $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ *from* $\mathcal{P}$, *and* $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})} \ln|\widehat{\mathcal{E}}|$.

4.2. PAC-Bayesian Generalization Bounds for ReED

To present the PAC-Bayesian generalization bounds for ReED, we make the following assumptions:

A.1 All activation functions ϕ, ρ, ψ in Definition 3.1 are Lipschitz-continuous with respect to the Euclidean norm of their input and output vectors, i.e., there exists a Lipschitz constant C_g such that $\|g(\mathbf{x}_1) - g(\mathbf{x}_2)\|_2 \leq C_g\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ for an activation function $g(\cdot)$ and any vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. For example, ReLU, LeakyReLU, Tanh, SoftPlus, Sigmoid, ArcTan, and Softsign are Lipschitz continuous functions with Lipschitz constant 1; these are 1-Lipschitz activation functions [54].

A.2 The training triplets in $\widehat{\mathcal{E}}$ are sampled from the finite full triplet set $\mathcal{E}$ without replacement.

A.3 Regarding the sizes of $\mathcal{E}$ and $\widehat{\mathcal{E}}$, we assume $|\mathcal{E}| \geq 40$ and $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$.

Let us consider a triplet classification model that uses the RAMP encoder in Definition 3.1 and the TD decoder in Definition 3.2. Using Theorem 4.3, we compute the generalization bound of this model in Theorem 4.4. Recall the graph diffusion matrix $\boldsymbol{S}_r^{(l)}$ in Definition 3.1. Given a fully observed knowledge graph G, the full triplet set $\mathcal{E}$ is finite. Therefore, for every possible training knowledge graph $\widehat{G}$ and the corresponding training triplet set $\widehat{\mathcal{E}}$ sampled from $\mathcal{E}$, the infinity norms of the graph diffusion matrices for $\widehat{G}$ exist per relation. Let k_r denote the maximum value of the infinity norms for all possible graph diffusion matrices among all layers for relation r. Note that k_r is independent of the choice of the training triplet set and dependent on the full triplet set.

Theorem 4.4 (Generalization Bound for ReED with Translational Distance Decoder)**.** *For any* $L \geq 0$, *let* $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \to \mathbb{R}^2$ *be a triplet classifier designed by the combination of the RAMP encoder with L-layers in Definition 3.1 and the TD decoder in Definition 3.2. Let* k_r *be the maximum of the infinity norms for all possible* $\boldsymbol{S}_r^{(l)}$ *in the RAMP encoder. Then, for any* $\delta, \gamma > 0$, *with probability at least* $1-\delta$ *over a training triplet set* $\widehat{\mathcal{E}}$ *(such that* $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$*) sampled without replacement from the full triplet set* $\mathcal{E}$, *for any* $\mathbf{w}$, *we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}}L^2\zeta_L^2 s^{2L} d\ln(N_{\mathbf{w}}d)}{\gamma^2} + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right)$$

where $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$, $N_{\mathbf{w}} = 2|\mathcal{R}|L + 6|\mathcal{R}| + 2L$, $\zeta_L = 2\tau^L\|\boldsymbol{X}_{\text{ent}}\|_2 + 2\kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{L-1}\tau^i\right) + \|\boldsymbol{X}_{\text{rel}}\|_2$, $\tau = C_\phi + \kappa$, $\kappa = C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} k_r$, $d = \max\left(\max_{0\leq l\leq L+1}(d_l), \max_{0\leq l\leq L+1}(d'_l)\right)$, $s_{L+1} = \max_{r,j}(\max(\|\overline{\boldsymbol{W}}_r^{\langle j\rangle}\|_F, \|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_F, \|\boldsymbol{V}_r^{\langle j\rangle}\|_F))$, $s_l = \max(\|\boldsymbol{W}_0^{(l)}\|_F, \|\boldsymbol{U}_0^{(l)}\|_F, \max_r\|\boldsymbol{W}_r^{(l)}\|_F, \max_r\|\boldsymbol{U}_r^{(l)}\|_F)$ *for* $l \in \{1, 2, \ldots, L\}$, *and* $s = \max_{1\leq l\leq L+1}(s_l)$.

In Theorem 4.5, we also compute the generalization bound of a model that uses the RAMP encoder in Definition 3.1 and the SM decoder in Definition 3.3.

Theorem 4.5 (Generalization Bound for ReED with Semantic Matching Decoder)**.** *For any* $L \geq 0$, *let* $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \to \mathbb{R}^2$ *be a triplet classifier designed by the combination of the RAMP encoder with L-layers in Definition 3.1 and the SM decoder in Definition 3.3. Let* k_r *be the maximum of the infinity norms for all possible* $\boldsymbol{S}_r^{(l)}$ *in the RAMP encoder. Then, for any* $\delta, \gamma > 0$, *with probability at least* $1-\delta$ *over a training triplet set* $\widehat{\mathcal{E}}$ *(such that* $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$*) sampled without replacement from the full triplet set* $\mathcal{E}$, *for any* $\mathbf{w}$, *we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}}L^2\eta_L^4 s^{4L} d\ln(N_{\mathbf{w}}d)}{\gamma^2} + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right)$$

where $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$, $N_{\mathbf{w}} = 2|\mathcal{R}|L + 2|\mathcal{R}| + 2L$, $\eta_L = \tau^L\|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\sum_{i=0}^{L-1}\tau^i$, $d = \max(\max_{0\leq l\leq L}(d_l), \max_{0\leq l\leq L}(d'_l))$, $\tau = C_\phi + \kappa$, $\kappa = C_\phi C_\rho C_\psi\sum_{r\in\mathcal{R}} k_r$, $s_{L+1} = \max_{r,j}\|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_F$, $s_l = \max(\|\boldsymbol{W}_0^{(l)}\|_F, \|\boldsymbol{U}_0^{(l)}\|_F, \max_r\|\boldsymbol{W}_r^{(l)}\|_F, \max_r\|\boldsymbol{U}_r^{(l)}\|_F)$ *for* $l \in \{1, 2, \ldots, L\}$, *and* $s = \max_{1\leq l\leq L+1}(s_l)$.

When we derive Theorem 4.4 and Theorem 4.5 from Theorem 4.3, we assume the prior distribution to be the Gaussian distribution with the zero mean and the standard deviation σ. We also assume that the perturbation follows the same Gaussian distribution as the prior distribution since the perturbation can follow any distribution, as indicated in Theorem 4.3. To derive the generalization bounds of ReED using Theorem 4.3, we need to find σ such that $\mathbb{P}(\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$ is satisfied, where $\ddot{\mathbf{w}}$ follows the Gaussian distribution with the zero mean and the standard deviation of σ. In this process, we express $\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty$ using our encoder (Definition 3.1) and decoder (Definition 3.2 or Definition 3.3). Note that σ should be independent of the learned parameters $\mathbf{w}$ since σ is the standard deviation of the prior distribution that should be independent of the training data. Thus, we use an approximation of the norm of $\mathbf{w}$ instead of its actual norm when computing σ; the actual norm of $\mathbf{w}$ is considered to be within a certain range from our approximation. Finally, we express the generalization bound in terms of the actual norm of $\mathbf{w}$ using the covering number arguments [4]. The full proofs of Theorem 4.4 and Theorem 4.5 are in Appendix D.

To prove Theorem 4.4 and Theorem 4.5, we follow some proof techniques of Liao et al. [24] considering standard graphs with a single relation. However, our proofs require more complex derivations than Liao et al. [24] because knowledge graphs have multiple relations. In our proofs, the interactions between the entities and the relations should be considered in the recurrence relation of an entity representation, leading to two-step unrolling of each recursion step. As a result, the generalization bounds are computed by considering (i) the norm of the difference between an unperturbed entity representation and a perturbed entity representation, (ii) the norm of the difference between an unperturbed relation representation and a perturbed relation representation, (iii) the norm of an entity representation, and (iv) the norm of a relation representation.

When we use one-hot encoding for $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$, the spectral norms of $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$ are one, and the maximum dimension becomes $d = |\mathcal{V}|$. Let us assume that ϕ, ρ, ψ are 1-Lipschitz activation functions. In this case, we can simplify $\zeta_L = 4\tau^L - 1$ in Theorem 4.4 and $\eta_L = 2\tau^L - 1$ in Theorem 4.5. In Corollary 4.6, we present a simplified form of our generalization bounds by leaving model-dependent terms and regarding the rest as a constant.

Corollary 4.6 (Simplified Form of the Generalization Bounds for ReED). *For any $L \geq 0$, let $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \rightarrow \mathbb{R}^2$ be a triplet classifier with the combination of the RAMP encoder with L-layers in Definition 3.1 and a decoder defined in Definition 3.2 or Definition 3.3. Let k_r be the maximum of the infinity norms for all possible $\boldsymbol{S}_r^{(l)}$ in the RAMP encoder. Then for any $\delta, \gamma > 0$, with probability at least $1 - \delta$ over a training triplet set $\widehat{\mathcal{E}}$ (such that $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$) drawn without replacement from the full triplet set $\mathcal{E}$, for any $\mathbf{w}$, we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \begin{cases} \mathcal{O}\left(L\tau^L s^L \sqrt{N_{\mathbf{w}} \ln N_{\mathbf{w}}}\right) & \text{(Translational Distance)} \\ \mathcal{O}\left(L\tau^{2L} s^{2L} \sqrt{N_{\mathbf{w}} \ln N_{\mathbf{w}}}\right) & \text{(Semantic Matching)} \end{cases}$$

where s is the maximum of the Frobenius norms of all learnable matrices and $\tau = 1 + \sum_{r \in \mathcal{R}} k_r$, and $N_{\mathbf{w}}$ is the total number of learnable matrices.

In Corollary 4.6, the bounds are largely affected by k_r which is the maximum infinity norm of all possible graph diffusion matrices across all layers for relation r. Recall two different ways of defining a graph diffusion matrix $\boldsymbol{S}_r^{(l)}$ discussed in Section 3.1: an adjacency matrix (corresponding to a sum aggregator) or a degree-normalized adjacency matrix (corresponding to a mean aggregator). Note that k_r becomes the maximum degree of an entity per relation when the sum aggregator is used, whereas k_r becomes at most one when the mean aggregator is used since each row is normalized. Thus, a mean aggregator can be a better option than a sum aggregator in reducing the generalization bounds. The total number of learnable matrices $N_{\mathbf{w}}$ is another critical factor: the generalization bounds decrease when the number of parameters is reduced. This can explain the effectiveness of the parameter-sharing strategies in Vashishth et al. [52] and the basis or block decomposition ideas in Schlichtkrull et al. [42]. On the other hand, the maximum of the Frobenius norms of learnable matrices s also critically affects the generalization bounds. Therefore, the generalization bounds decrease when the weight matrices or the entity/relation representations are normalized. This observation can provide theoretical justification for weight normalization adapted in Oono & Suzuki [39] and normalization of entity representations used in Bordes et al. [8]. In Corollary 4.6, τs is usually greater than one because s is typically not less than one [18]. Thus, the generalization bounds sharply increase when the number of encoder layers L increases.

5. Experiments

We conduct experiments on three real-world knowledge graphs: FB15K237 [47], CoDEx-M [40], and UMLS-43 [7; 27]. Note that FB15K237 and CoDEx-M are well-known knowledge graph benchmarks extracted from commonly used knowledge bases, Freebase and Wikidata, respectively, and UMLS-43 is another benchmark extracted from a popular biomedical knowledge base, UMLS. On all datasets, we create $\widehat{\mathcal{E}}$ randomly sampled from $\mathcal{E}$ without replacement with the sampling probability of 0.8. In the ReED framework, we use the RAMP encoder in Definition 3.1 with L layers and use either the TD or SM decoder (Definition 3.2 and Definition 3.3). In the RAMP encoder, we use $\rho = \psi =$ identity and $\phi =$ LeakyReLU. We use one-hot encoding for $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$. More details about the datasets and the settings are in Appendix E.

We measure the generalization errors on real-world datasets, where a generalization error is the actual difference between the expected and empirical losses empirically observed in a particular experiment; the generalization bound is the theoretical upper bound of these generalization errors. In Corollary 4.6, among the factors that affect the generalization bounds, we empirically measure the effects of the following three factors: (i) whether a model uses a mean aggregator or a sum aggregator, which affects k_r, (ii) the Frobenius norms of the learnable matrices s, and (iii) the number of layers L in the RAMP encoder[2]. We compare the generalization errors by varying one of these three factors while the other two factors are controlled. Note that s is defined as the maximum of Frobenius norms of all learnable matrices. Indeed, to control the effect of the norms of these

[2]Note that $N_{\mathbf{w}}$ is proportional to L since $N_{\mathbf{w}} = \mathcal{O}(|\mathcal{R}|L)$ as indicated in Theorem 4.4 and Theorem 4.5.

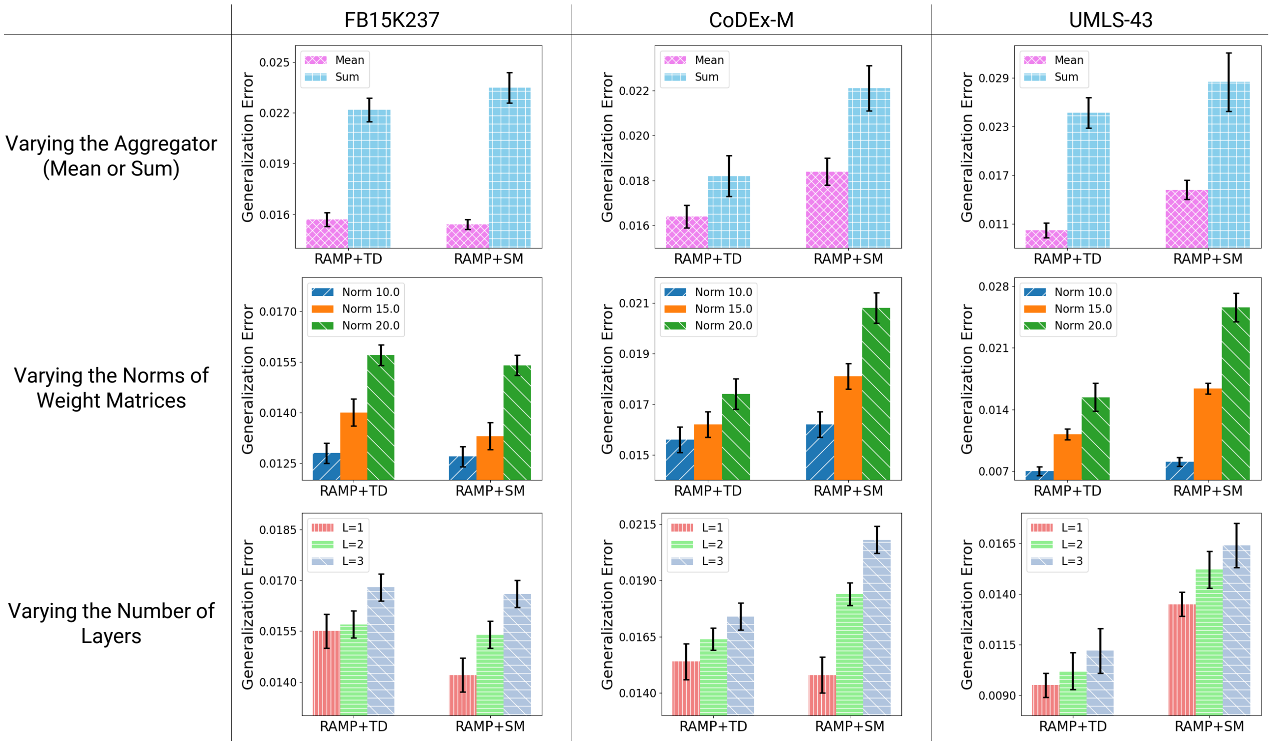

Figure 2: Generalization Errors of ReED according to different aggregators, norms of the weight matrices, and numbers of layers in the RAMP encoder. In ReED, two different triplet classification decoders, TD or SM, are used. The changing trends in generalization errors according to the three different factors align with the theoretical findings in Corollary 4.6.

matrices more precisely, the Frobenius norm of each weight matrix should be fixed to be s as described in the proof of our generalization bounds in Appendix D. Thus, we normalize each weight matrix after each backpropagation step. We use $s \in \{10.0, 15.0, 20.0\}$ and $L \in \{1, 2, 3\}$ for all datasets. Figure 2 shows the generalization errors of ReED depending on the decoder: RAMP+TD and RAMP+SM. We repeat all experiments 10 times and visualize the mean and the standard deviation. Across all datasets and all models, the mean aggregator shows lower generalization errors than the sum aggregator. Also, the generalization errors increase as s and L increase. These empirical observations are aligned with our theoretical findings in Corollary 4.6. Even though the theoretical generalization bound indicates the upper limit of the possible generalization errors, Figure 2 shows that the critical factors explaining the generalization bounds also affect an actual generalization error.

We conduct additional experiments using the initial features of entities and relations instead of using one-hot encoding for $\boldsymbol{X}_{\text{ent}}$ and $\boldsymbol{X}_{\text{rel}}$ on FB15K237 to observe the generalization errors by varying d. We extract the initial features by feeding the textual descriptions of entities and relations to BERT [12] and reduce the dimension of the extracted features to 32 using PCA; we use the resulting features as $\mathbf{X}_{\text{ent}}$ and $\mathbf{X}_{\text{rel}}$. Then, we calculate the generalization errors of ReED according to different maximum dimensions d while fixing the other factors (e.g., the aggregator, the norms of weight matrices, and the number of layers). Figure 3 shows the generalization errors of ReED with different $d(= d_1 = d_2 = \cdots = d_L = d_{L+1} = d'_1 = d'_2 = \cdots = d'_L = d'_{L+1}) \in \{64, 96, 128\}$. We observe that the generalization errors increase as d increases, which aligns with the expected tendency.

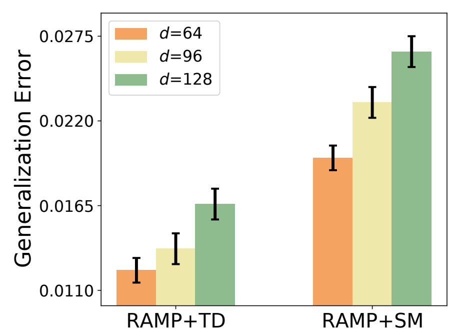

Figure 3: Generalization Errors of ReED on FB15K237 according to different maximum dimensions d.

6. Related Work and Discussion

Regarding the generalization ability of knowledge graph embedding, Kuželka & Wang [21] have computed the expected number of incorrect predictions made by knowledge graph embedding methods, which differs from a standard generalization bound defined by the difference between the expected and the empirical errors; their work is neither applicable to GNN-based models nor a margin loss.

While there have been some studies about the generalization bounds for GNNs [24; 18; 29], they have considered graph classification tasks on standard graphs with a single relation. For example, Liao et al. [24] assume that graphs are i.i.d. samples drawn from some unknown infinite distribution. Also, Ju et al. [18] consider a twice-differentiable loss function to compute Hessian-based bounds, and Maskey et al. [29] apply MPNNs on the underlying continuous space from which graphs are sampled. Our work and these previous works [24; 18; 29] are significantly distinct in that (i) we deal with knowledge graphs having multiple relations, (ii) our target task is a triplet classification, (iii) we assume a finite full set since the triplets are finite given a fixed knowledge graph while the previous studies [24; 18; 29] assume that graphs are sampled from an infinite space.

Different PAC-Bayes approaches have been explored in various perspectives. For example, Guedj [16] provides a survey about the PAC-Bayes framework, including the extention of the KL divergence to f-divergence for expressing a more general divergence class in computing PAC-Bayes bounds. Also, Alquier [1] provides a recent survey about various tight PAC-Bayes bounds in varied settings. Though our study focuses on the traditional KL divergence and considers the transductive PAC-Bayesian approach, we expect our work to be extended to a broader class of divergences or information-theoretic approaches.

7. Conclusion and Future Work

To comprehensively analyze the generalization bounds for KGRL, we propose a generic framework, ReED, that can subsume many existing KGRL methods. We prove the PAC-Bayesian generalization bounds for ReED having two different triplet classification decoders. Our analysis provides theoretical evidence for the benefits of the parameter-sharing and weight normalization schemes and the advantage of a mean aggregator over a sum aggregator within a neural encoder in reducing the generalization bounds in KGRL.

We note that the ReED framework cannot exhaustively cover all existing KGRL methods. Specifically, the graph attention networks [33; 53; 9] are hard to consider in ReED with the current form. Extending ReED to the attention mechanisms is one of our future works. Also, we plan to investigate the relationships between the generalization ability and the expressivity [32] in KGRL based on our findings in the generalization bounds of KGRL.

Impact Statement

Most of our contributions in this paper are theoretical, and our work aims to advance the field of Machine Learning at a fundamental level. Considering that knowledge graphs are widely utilized in information retrieval (e.g., Google Knowledge Graph), a societal consequence of our work is to improve the retrieval performance by providing theoretical insights for KGRL methods. Our findings and their practical implications can guide the desirable designs of future KGRL methods. Generally speaking, our generalization bounds indicate that reducing the number of learnable parameters, the norms of weight matrices, and the maximum infinity norm of the graph diffusion matrices is beneficial to decreasing the generalization bounds.

Acknowledgements

This research was supported by an NRF grant funded by MSIT 2022R1A2C4001594 (Extendable Graph Representation Learning) and an IITP grant funded by MSIT 2022-0-00369 (Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

References

[1] Alquier, P. User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

[2] Barcelo, P., Galkin, M., Morris, C., and Orth, M. R. Weisfeiler and Leman go relational. In *Proceedings of the 1st Learning on Graphs Conference*, pp. 46:1–46:26, 2022.

[3] Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

[4] Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 6240–6249, 2017.

[5] Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[6] Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F.

PAC-Bayesian theory for transductive learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pp. 105–113, 2014.

[7] Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

[8] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pp. 2787–2795, 2013.

[9] Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[10] Chao, L., He, J., Wang, T., and Chu, W. PairRE: Knowledge graph embeddings via paired relation vectors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4360–4369, 2021.

[11] Chung, C., Lee, J., and Whang, J. J. Representation learning on hyper-relational and numeric knowledge graphs with transformers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 310–322, 2023.

[12] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[13] Esser, P. M., Vankadara, L. C., and Ghoshdastidar, D. Learning theory can (sometimes) explain generalisation in graph neural networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 27043–27056, 2021.

[14] Garg, V. K., Jegelka, S., and Jaakkola, T. Generalization and representational limits of graph neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3419–3430, 2020.

[15] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

[16] Guedj, B. A primer on pac-bayesian learning. In *arXiv preprint*, 2019.

[17] Huang, X., Orth, M. R., Ceylan, İ. İ., and Barceló, P. A theory of link prediction via relational Weisfeiler-Leman on knowledge graphs. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[18] Ju, H., Li, D., Sharma, A., and Zhang, H. R. Generalization in graph neural networks: Improved PAC-Bayesian bounds on graph diffusion. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 6314–6341, 2023.

[19] Kazemi, S. M. and Poole, D. SimplE embedding for link prediction in knowledge graphs. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pp. 4284–4295, 2018.

[20] Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[21] Kuželka, O. and Wang, Y. Generalization bounds for knowledge graph embedding (trained by maximum likelihood). NeurIPS 2019 Workshop on Machine Learning with Guarantees, 2019.

[22] Lee, J., Chung, C., Lee, H., Jo, S., and Whang, J. J. VISTA: Visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7314–7328, 2023.

[23] Lee, J., Chung, C., and Whang, J. J. InGram: Inductive knowledge graph embedding via relation graphs. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 18796–18809, 2023.

[24] Liao, R., Urtasun, R., and Zemel, R. A PAC-Bayesian approach to generalization bounds for graph neural networks. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[25] Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2181–2187, 2015.

[26] Liu, H., Wu, Y., and Yang, Y. Analogical inference for multi-relational embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2168–2178, 2017.

[27] Lloyd, O., Liu, Y., and Gaunt, T. R. Assessing the effects of hyperparameters on knowledge graph embedding quality. *Journal of Big Data*, 10(1):59, 2023.

[28] Ma, J., Deng, J., and Mei, Q. Subgroup generalization and fairness of graph neural networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 1048–1061, 2021.

[29] Maskey, S., Levie, R., Lee, Y., and Kutyniok, G. Generalization analysis of message passing neural networks on large random graphs. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, pp. 4805–4817, 2022.

[30] McAllester, D. Simplified PAC-Bayesian margin bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, pp. 203–215, 2003.

[31] McAllester, D. A. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 230–234, 1998.

[32] Morris, C., Geerts, F., Tönshoff, J., and Grohe, M. WL meet VC. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 25275–25302, 2023.

[33] Nathani, D., Chauhan, J., Sharma, C., and Kaul, M. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4710–4723, 2019.

[34] Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 5947–5956, 2017.

[35] Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[36] Nickel, M., Tresp, V., and Kriegel, H.-P. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 809–816, 2011.

[37] Nickel, M., Rosasco, L., and Poggio, T. Holographic embeddings of knowledge graphs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 1955–1961, 2016.

[38] Oono, K. and Suzuki, T. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, pp. 18917–18930, 2020.

[39] Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[40] Safavi, T. and Koutra, D. CoDEx: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8328–8350, 2020.

[41] Scarselli, F., Tsoi, A. C., and Hagenbuchner, M. The Vapnik-Chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

[42] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *Proceedings of the 15th Extended Semantic Web Conference*, pp. 593–607, 2018.

[43] Shang, C., Tang, Y., Huang, J., Bi, J., He, X., and Zhou, B. End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 3060–3067, 2019.

[44] Socher, R., Chen, D., Manning, C. D., and Ng, A. Y. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pp. 926–934, 2013.

[45] Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[46] Teru, K. K., Denis, E. G., and Hamilton, W. L. Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9448–9457, 2020.

[47] Toutanova, K. and Chen, D. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

[48] Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[49] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2071–2080, 2016.

[50] Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[51] Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[52] Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. Composition-based multi-relational graph convolutional networks. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[53] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[54] Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pp. 3835–3844, 2018.

[55] Wang, Q., Mao, Z., Wang, B., and Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[56] Wang, Z., Zhang, J., Feng, J., and Chen, Z. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pp. 1112–1119, 2014.

[57] Weisfeiler, B. and Lehman, A. A. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia*, 2(9):12–16, 1968.

[58] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

[59] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[60] Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[61] Zhang, S., Tay, Y., Yao, L., and Liu, Q. Quaternion knowledge graph embeddings. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pp. 2735–2745, 2019.

[62] Zhou, Y., Kutyniok, G., and Ribeiro, B. OOD link prediction generalization capabilities of message-passing GNNs in larger test graphs. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, pp. 20257–20272, 2022.

A. Basic Notation

In Table 3, we provide a concise overview of the notation used throughout the paper. Any other notation not listed in Table 3 is clarified and detailed within the context.

Table 3: Overview of basic notation

Symbol	Meaning
$[\cdot \quad \cdot], \begin{bmatrix}\cdot \\ \cdot\end{bmatrix}$	a horizontal/vertical concatenation
$\lvert \cdot \rvert$	the size of a set or the absolute value of a scalar value
$\lVert \cdot \rVert_2, \lVert \cdot \rVert_\infty, \lVert \cdot \rVert_F$	the spectral(Euclidean)/infinity/Frobenius norm of a matrix(vector)
$\text{diag}(\mathbf{x})$	a diagonal matrix whose diagonal is defined by the vector $\mathbf{x}$
$\boldsymbol{I}_{n\times n}$	an identity matrix of size $n \times n$
$\mathbf{0}_{m\times n}$	an all-zero matrix of size $m \times n$
$\mathbf{0}_n$	an all-zero vector of size n
$\mathbb{1}[\cdot]$	an indicator function
$\mathbb{P}[\cdot], \mathbb{E}[\cdot]$	the probability/expectation
$\phi(\cdot), \rho(\cdot), \psi(\cdot)$	Lipschitz-continuous activation functions
C_ϕ, C_ρ, C_ψ	Lipschitz constants of ϕ, ρ, ψ
$\boldsymbol{M}^\top$	the transpose of a matrix $\boldsymbol{M}$
$G, \widehat{G}$	a fully observed/training knowledge graph
$\mathcal{V}, \mathcal{R}$	a set of entities/relations
$\mathcal{E}, \widehat{\mathcal{E}}$	a full triplet set/training triplet set
$\boldsymbol{X}_{\text{ent}}, \boldsymbol{X}_{\text{rel}}$	a matrix of entity/relation features
d_0, d'_0	the dimension of the initial feature vector of an entity/relation
d_l, d'_l	the dimension of an entity/relation representation at the l-th layer
L	the total number of layers in the RAMP encoder
y_{hrt}	the ground-truth label of a triplet $(h, r, t) \in \mathcal{E}$
$\boldsymbol{H}, \boldsymbol{R}$	a matrix of entity/relation representations
$\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{V}$	the learnable projection matrices
$\boldsymbol{S}_r$	the graph diffusion matrix of $\widehat{G}$ for relation $r \in \mathcal{R}$
$D_{KL}(\mathcal{Q}\Vert\mathcal{P})$	KL-divergence of $\mathcal{Q}$ from $\mathcal{P}$
$\mathcal{P}, \mathcal{Q}$	a prior/posterior distribution on a hypothesis space $\mathcal{H}$
$\ln(\cdot)$	the natural logarithm

B. Interpreting ReED as a Generalization of Existing KGRL Methods

Our ReED framework consists of the RAMP encoder and a triplet classification decoder as described in Section 3. We provide the details about how our RAMP encoder and two types of triplet classification decoders (i.e., translational distance decoder and semantic matching decoder) can express diverse KGRL methods.

B.1. Representing Existing KGRL Encoders Using RAMP Encoder

In Section 3.1, we define the RAMP encoder in Definition 3.1 and show that several well-known GNN-based KGRL encoders can be considered as special cases of our RAMP encoder. For example, R-GCN [42], WGCN [43], and CompGCN [52] can be represented using the RAMP encoder by appropriately setting the activation functions ϕ, ρ, ψ, the projection matrices $\boldsymbol{W}_r^{(l)}$ and $\boldsymbol{U}_r^{(l)}$, and the graph diffusion matrices $\boldsymbol{S}_r^{(l)}$ in Definition 3.1, as shown in Table 1.

Note that CompGCN has three different variations depending on the composition operator: subtraction (Sub), multiplication (Mult), and circular correlation (Corr). We detail how CompGCN (Corr) can be represented using the RAMP encoder here; the others are all described in the main paper. Give an entity representation $\boldsymbol{H}^{(l)}[v,:]$ and a relation representation $\boldsymbol{R}^{(l)}[r,:]$ where both have the dimension of d, the circular correlation $\star$ is defined by $(\boldsymbol{H}^{(l)}[v,:] \star \boldsymbol{R}^{(l)}[r,:])[k] :=$

$\sum_{i=0}^{d-1} \boldsymbol{H}^{(l)}[v,i]\boldsymbol{R}^{(l)}[r,\ (k+i) \bmod d]$ where $k \in \{0,1,...,d-1\}$. Let us define $\boldsymbol{C}_r^{(l)}$ as

$$\boldsymbol{C}_r^{(l)} := \begin{bmatrix} \boldsymbol{R}^{(l)}[r,0] & \boldsymbol{R}^{(l)}[r,1] & \cdots & \boldsymbol{R}^{(l)}[r,d'_l-1] \\ \boldsymbol{R}^{(l)}[r,1] & \boldsymbol{R}^{(l)}[r,2] & \cdots & \boldsymbol{R}^{(l)}[r,0] \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}^{(l)}[r,d'_l-1] & \boldsymbol{R}^{(l)}[r,0] & \cdots & \boldsymbol{R}^{(l)}[r,d'_l-2] \end{bmatrix}.$$

By setting ϕ to Tanh, ρ, ψ to the identity functions, $\boldsymbol{W}_r^{(l)} := \boldsymbol{C}_r^{(l-1)}\boldsymbol{W}_{\lambda(r)}^{(l)}$, $\boldsymbol{U}_r^{(l)} := \boldsymbol{0}_{d'_{l-1}\times d_l}$, and $\boldsymbol{S}_r^{(l)}[v,:] := \boldsymbol{A}_r[v,:]$, we can represent CompGCN (Corr) using our RAMP encoder. Note that the dimensions of the entity and relation representations should be the same for all layers in CompGCN.

B.2. Translational Distance Decoder and Existing KGRL Methods

We define our translational distance (TD) decoder in Definition 3.2, which includes three learnable projection matrices: $\overline{\boldsymbol{W}}_r^{\langle j\rangle}$, $\boldsymbol{V}_r^{\langle j\rangle}$ and $\overline{\boldsymbol{U}}_r^{\langle j\rangle}$. By appropriately defining these matrices, our TD decoder can express five different knowledge graph embedding methods: TransE [8], TransH [56], TransR [25], RotatE [45], and PairRE [10]. Since we already described TransR and RotatE in the main paper, we describe the other three methods here. Recall that we introduce two learnable matrices, $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle} \in \mathbb{R}^{d_L\times \bar{d}}$ and $\boldsymbol{T}_{\text{rel}}^{\langle j\rangle} \in \mathbb{R}^{d'_L\times \bar{d}'}$, which are only needed for specializing our decoder to simulate an existing shallow-architecture knowledge graph embedding model, as described in Section 3.2.

TransE [8] For each triplet, TransE assumes that the heady entity representation is translated by the relation representation, and the resulting vector should be placed close to the tail entity representation. Our Definition 3.2 can trivially express TransE by setting

$$\overline{\boldsymbol{W}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle} \qquad \boldsymbol{V}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle} \qquad \overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{rel}}^{\langle j\rangle}$$

where $\bar{d} = \bar{d}' = d_{L+1}$. Note that TransE has a constraint that $\|\boldsymbol{H}^{(L)}[v,:]\boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\|_2 = 1, \forall v \in \mathcal{V}$ for $j \in \{0,1\}$.

TransH [56] In TransH, each entity representation vector is projected onto a relation-specific hyperplane, where a projected head entity is translated by the relation representation, and the resulting vector is assumed to be placed close to the projected tail entity representation. For relation r, let $\mathbf{f}_r^{\langle j\rangle} \in \mathbb{R}^{\bar{d}}$ denote the unit normal vector of the hyperplane for r. Note that $\|\mathbf{f}_r^{\langle j\rangle}\|_2 = 1$ for $j \in \{0,1\}$ and $\bar{d} = \bar{d}' = d_{L+1}$. By Setting

$$\overline{\boldsymbol{W}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}(\boldsymbol{I}_{\bar{d}\times\bar{d}} - \mathbf{f}_r^{\langle j\rangle\top}\mathbf{f}_r^{\langle j\rangle}) \qquad \boldsymbol{V}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}(\boldsymbol{I}_{\bar{d}\times\bar{d}} - \mathbf{f}_r^{\langle j\rangle\top}\mathbf{f}_r^{\langle j\rangle}) \qquad \overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{rel}}^{\langle j\rangle},$$

our TD decoder can represent TransH.

PairRE [10] In PairRE, a relation representation comprises two parts: representations for the head and the tail. The head and tail entity representations are translated by the relation representation corresponding to the head and tail part, respectively. The translated entity representations are assumed to be close to each other. Let $\mathfrak{f}_r^{\langle j\rangle} \in \mathbb{R}^{\bar{d}'}$ denote the representation of r for translating head entities and $\mathring{\mathfrak{f}}_r^{\langle j\rangle} \in \mathbb{R}^{\bar{d}'}$ denote the representation of r for translating tail entities. Let us define $\mathfrak{F}_r^{\langle j\rangle} = \text{diag}\left(\mathfrak{f}_r^{\langle j\rangle}\right)$ and $\mathring{\mathfrak{F}}_r^{\langle j\rangle} = \text{diag}\left(\mathring{\mathfrak{f}}_r^{\langle j\rangle}\right)$. It is assumed that the entity representations are on the unit circle, i.e., $\|\boldsymbol{H}^{(L)}[v,:]\boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\|_2 = 1, \forall v \in \mathcal{V}$ for $j \in \{0,1\}$. Given $\bar{d} = \bar{d}' = d_{L+1}$, we can represent PairRE by setting

$$\overline{\boldsymbol{W}}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\mathfrak{F}_r^{\langle j\rangle} \qquad \boldsymbol{V}_r^{\langle j\rangle} := \boldsymbol{T}_{\text{ent}}^{\langle j\rangle}\mathring{\mathfrak{F}}_r^{\langle j\rangle} \qquad \overline{\boldsymbol{U}}_r^{\langle j\rangle} := \boldsymbol{0}_{d'_L\times d_{L+1}}.$$

B.3. Semantic Matching Decoder and Existing KGRL Methods

Our semantic matching (SM) decoder in Definition 3.3 can represent seven different knowledge graph embedding methods: RESCAL [36], DistMult [60], HolE [37], ComplEx [49], ANALOGY [26], SimplE [19], and QuatE [61]. To show our SM decoder can be specialized to express these existing methods, we introduce two learnable matrices, $\boldsymbol{T}_{\text{ent}}^{\langle j\rangle} \in \mathbb{R}^{d_L\times\bar{d}}$ and $\boldsymbol{T}_{\text{rel}}^{\langle j\rangle} \in \mathbb{R}^{d'_L\times\bar{d}'}$. Note that we need to only define $\overline{\boldsymbol{U}}_r^{\langle j\rangle}$ in Definition 3.3 to simulate the existing methods. We omit DistMult and ANALOGY here since they are described in the main paper.

RESCAL [36] By defining a relation-specific projection matrix $\boldsymbol{B}_r^{\langle j \rangle} \in \mathbb{R}^{\bar{d}\times\bar{d}}$, we can easily represent RESCAL by setting

$$\overline{\boldsymbol{U}}_r^{\langle j \rangle} \coloneqq \boldsymbol{T}_{\text{ent}}^{\langle j \rangle}\boldsymbol{B}_r^{\langle j \rangle}{\boldsymbol{T}_{\text{ent}}^{\langle j \rangle}}^\top.$$

HolE [37] and ComplEx [49] It has been known that HolE and ComplEx are special cases of ANALOGY [26]. Since our SM decoder can represent ANALOGY as described in Section 3.2, our decoder can also represent HolE and ComplEx.

SimplE [19] In SimplE, an entity representation is divided into two parts: the first part is the representation when the entity appears as a head entity, and the second part is the representation when the entity appears as a tail entity. For each relation r, its inverse relation r^{-1} is also considered. As a result, a relation representation is also divided into two parts: representations for r and r^{-1}. Given a triplet (h, r, t), SimplE calculates its score by averaging the scores of (h, r, t) and (t, r^{-1}, h). Given $\bar{d} = \bar{d}'$, we can formulate SimplE by setting

$$\overline{\boldsymbol{U}}_r^{\langle j \rangle} \coloneqq \frac{1}{2}\boldsymbol{T}_{\text{ent}}^{\langle j \rangle}\left(\text{diag}\Big(\big(\boldsymbol{R}^{(L)}\boldsymbol{T}_{\text{rel}}^{\langle j \rangle}\big)[r,:]\Big)\begin{bmatrix}\mathbf{0}_{\bar{d}'/2,\bar{d}'/2} & \boldsymbol{I}_{\bar{d}'/2,\bar{d}'/2}\\ \boldsymbol{I}_{\bar{d}'/2,\bar{d}'/2} & \mathbf{0}_{\bar{d}'/2,\bar{d}'/2}\end{bmatrix}\right){\boldsymbol{T}_{\text{ent}}^{\langle j \rangle}}^\top.$$

QuatE [61] In QuatE, each representation vector is represented in the quaternion space. Let us denote a representation vector of r for $\langle j \rangle$ in the quaternion space as a quaternion of real numbers such that $\begin{bmatrix}\mathfrak{a}_r^{\langle j \rangle} & \mathfrak{b}_r^{\langle j \rangle} & \mathfrak{c}_r^{\langle j \rangle} & \mathfrak{d}_r^{\langle j \rangle}\end{bmatrix}$ where $\mathfrak{a}_r^{\langle j \rangle}$ is the real part and $\mathfrak{b}_r^{\langle j \rangle}$, $\mathfrak{c}_r^{\langle j \rangle}$, and $\mathfrak{d}_r^{\langle j \rangle}$ correspond to the imaginary coefficients. It is assumed that $(\mathfrak{a}_r^{\langle j \rangle}[i])^2 + (\mathfrak{b}_r^{\langle j \rangle}[i])^2 + (\mathfrak{c}_r^{\langle j \rangle}[i])^2 + (\mathfrak{d}_r^{\langle j \rangle}[i])^2 = 1$ for $i \in \{0, 1, \cdots, \frac{\bar{d}'}{4}-1\}$ and $j \in \{0, 1\}$. Let us define

$$\mathfrak{A}_r^{\langle j \rangle} \coloneqq \text{diag}\Big(\mathfrak{a}_r^{\langle j \rangle}\Big) \in \mathbb{R}^{\bar{d}'/4\times\bar{d}'/4}, \mathfrak{B}_r^{\langle j \rangle} \coloneqq \text{diag}\Big(\mathfrak{b}_r^{\langle j \rangle}\Big) \in \mathbb{R}^{\bar{d}'/4\times\bar{d}'/4}$$

$$\mathfrak{C}_r^{\langle j \rangle} \coloneqq \text{diag}\Big(\mathfrak{c}_r^{\langle j \rangle}\Big) \in \mathbb{R}^{\bar{d}'/4\times\bar{d}'/4}, \mathfrak{D}_r^{\langle j \rangle} \coloneqq \text{diag}\Big(\mathfrak{d}_r^{\langle j \rangle}\Big) \in \mathbb{R}^{\bar{d}'/4\times\bar{d}'/4}.$$

Given $\bar{d} = \bar{d}'$, we can express QuatE using our SM decoder by setting

$$\overline{\boldsymbol{U}}_r^{\langle j \rangle} \coloneqq \boldsymbol{T}_{\text{ent}}^{\langle j \rangle}\begin{bmatrix}\mathfrak{A}_r^{\langle j \rangle} & \mathfrak{B}_r^{\langle j \rangle} & \mathfrak{C}_r^{\langle j \rangle} & \mathfrak{D}_r^{\langle j \rangle}\\ -\mathfrak{B}_r^{\langle j \rangle} & \mathfrak{A}_r^{\langle j \rangle} & -\mathfrak{D}_r^{\langle j \rangle} & \mathfrak{C}_r^{\langle j \rangle}\\ -\mathfrak{C}_r^{\langle j \rangle} & \mathfrak{D}_r^{\langle j \rangle} & \mathfrak{A}_r^{\langle j \rangle} & -\mathfrak{B}_r^{\langle j \rangle}\\ -\mathfrak{D}_r^{\langle j \rangle} & -\mathfrak{C}_r^{\langle j \rangle} & \mathfrak{B}_r^{\langle j \rangle} & \mathfrak{A}_r^{\langle j \rangle}\end{bmatrix}{\boldsymbol{T}_{\text{ent}}^{\langle j \rangle}}^\top.$$

C. Proof of Theorem 4.3

In Section 4.1, we present our Theorem 4.3 which states the transductive PAC-Bayesian generalization bound for a deterministic triplet classifier. We derive Theorem 4.3 from the following Lemma C.1 which is originally presented as Corollary 7 in Bégin et al. [6] where a transductive PAC-Bayesian generalization bound is presented for a stochastic model. We paraphrase the original version to customize it to our problem setting.

Lemma C.1 (Bégin et al. [6], Corollary 7)**.** *For any full triplet set $\mathcal{E}$ having size $|\mathcal{E}| \geq 40$, for any stochastic triplet classifier $\widetilde{f}$ following a posterior distribution $\mathcal{Q}$ on a hypothesis space $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any $\gamma, \delta > 0$, with probability at least $1-\delta$, over the choice of a training triplet set $\widehat{\mathcal{E}}$ (such that $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$) drawn without replacement from the full triplet set $\mathcal{E}$, we have*

$$\mathcal{L}_{\gamma,\mathcal{E}}\left(\widetilde{f}\right) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}\left(\widetilde{f}\right) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[D_{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|)}{\delta}\right]}$$

where $\mathcal{L}_{\gamma,\mathcal{Z}}\left(\widetilde{f}\right) \coloneqq \mathbb{E}_{\widetilde{\mathbf{w}}\sim\mathcal{Q}}\big[\mathcal{L}_{\gamma,\mathcal{Z}}(f_{\widetilde{\mathbf{w}}})\big]$, $f_{\widetilde{\mathbf{w}}}$ is a deterministic triplet classifier with parameters $\widetilde{\mathbf{w}}$, $\mathcal{L}_{\gamma,\mathcal{Z}}(f_{\widetilde{\mathbf{w}}})$ is defined in Definition 4.1, $D_{KL}(\mathcal{Q}\|\mathcal{P})$ is the KL-divergence of $\mathcal{Q}$ from $\mathcal{P}$, and $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$.

Note that the prior distribution $\mathcal{P}$ is independent of the training triplets. While Lemma C.1 considers a stochastic model, we need to consider a deterministic model because the ReED framework results in a deterministic triplet classifier. Given a deterministic triplet classifier $f_{\mathbf{w}}$ with the fixed parameters $\mathbf{w}$, we add random perturbations $\ddot{\mathbf{w}}$ to $\mathbf{w}$ to simulate a stochastic triplet classifier so that Lemma C.1 can be extended to Theorem 4.3.

Theorem 4.3 (Transductive PAC-Bayesian Generalization Bound for a Deterministic Triplet Classifier)**.** *Let $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \to \mathbb{R}^2$ be a deterministic triplet classifier with parameters $\mathbf{w}$, and $\mathcal{P}$ be any prior distribution on $\mathbf{w}$. Let us consider the finite full triplet set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$. We construct a posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ by adding any random perturbation $\ddot{\mathbf{w}}$ to $\mathbf{w}$ such that $\mathbb{P}(\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$. Then, for any $\gamma, \delta > 0$, with probability $1-\delta$ over the choice of a training triplet set $\widehat{\mathcal{E}}$ drawn from the full triplet set $\mathcal{E}$ (such that $20 \le |\widehat{\mathcal{E}}| \le |\mathcal{E}| - 20$ and $|\mathcal{E}| \ge 40$) without replacement, for any $\mathbf{w}$, we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \le \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1 - \frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln\frac{4\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|)}{\delta}\right]}$$

where $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ is defined in Definition 4.1, $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ is defined in Definition 4.2, $D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P})$ is the KL-divergence of $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ from $\mathcal{P}$, and $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1 - \frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$.

Proof. Let $\mathcal{H}$ denote the hypothesis space of a triplet classifier $f_{\mathbf{w}}$. Note that the posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ is a distribution over $\mathcal{H}$. Let $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}(\widetilde{\mathbf{w}})$ be the probability density function indicating the probability of $\widetilde{\mathbf{w}}$ being drawn from $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$. Also, let $\mathcal{C}$ be a set of perturbed parameters $\widetilde{\mathbf{w}}$ such that $\mathcal{C} \coloneqq \left\{\widetilde{\mathbf{w}} \in \mathcal{H} \middle| \max_{(h,r,t)\in\mathcal{E}} \|f_{\widetilde{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}\right\} \subset \mathcal{H}$. If we define $p \coloneqq \mathbb{P}_{\widetilde{\mathbf{w}}\sim\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}(\widetilde{\mathbf{w}} \in \mathcal{C})$, then $p > \frac{1}{2}$ by our assumption.

We divide $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ into two distributions $\grave{\mathcal{Q}}$ and $\acute{\mathcal{Q}}$ where the former has a non-zero value for $\widetilde{\mathbf{w}} \in \mathcal{C}$ and the latter has a non-zero value for $\widetilde{\mathbf{w}} \in \mathcal{H} \setminus \mathcal{C}$. Specifically, $\grave{\mathcal{Q}}(\widetilde{\mathbf{w}})$ and $\acute{\mathcal{Q}}(\widetilde{\mathbf{w}})$ are defined as

$$\grave{\mathcal{Q}}(\widetilde{\mathbf{w}}) = \begin{cases} \frac{1}{p}\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}(\widetilde{\mathbf{w}}) & \widetilde{\mathbf{w}} \in \mathcal{C} \\ 0 & \widetilde{\mathbf{w}} \in \mathcal{H}\setminus\mathcal{C} \end{cases}, \quad \acute{\mathcal{Q}}(\widetilde{\mathbf{w}}) = \begin{cases} 0 & \widetilde{\mathbf{w}} \in \mathcal{C} \\ \frac{1}{1-p}\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}(\widetilde{\mathbf{w}}) & \widetilde{\mathbf{w}} \in \mathcal{H}\setminus\mathcal{C} \end{cases}$$

For any $(h,r,t) \in \mathcal{E}$ and $\widetilde{\mathbf{w}} \sim \grave{\mathcal{Q}}$, we have

$$\begin{aligned}
&|(f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}]) - (f_{\mathbf{w}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}])| \\
=&|(f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[y_{hrt}]) - (f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}])| \\
\le&|f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[y_{hrt}]| + |f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]| \\
\le&\max_{(h,r,t)\in\mathcal{E}}(|f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[y_{hrt}]| + |f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]|) \\
<&\frac{\gamma}{4} + \frac{\gamma}{4} = \frac{\gamma}{2} \quad \text{(sub-additivity, } \widetilde{\mathbf{w}} \in \mathcal{C}\text{)}
\end{aligned}$$

Then, we have

$$\begin{aligned}
&f_{\mathbf{w}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}] \le 0 \Rightarrow f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] \le \frac{\gamma}{2} \\
&f_{\mathbf{w}}(h,r,t)[y_{hrt}] \le f_{\mathbf{w}}(h,r,t)[1-y_{hrt}] \Rightarrow f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] \le \frac{\gamma}{2} + f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}]
\end{aligned}$$

which indicates that

$$\mathbb{1}\left[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \le f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]\right] \le \mathbb{1}\left[f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] \le \frac{\gamma}{2} + f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}]\right]$$

This leads to $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \le \mathcal{L}_{\frac{\gamma}{2},\mathcal{E}}(f_{\widetilde{\mathbf{w}}})$ for any $\widetilde{\mathbf{w}} \sim \grave{\mathcal{Q}}$, meaning that

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \le \mathbb{E}_{\widetilde{\mathbf{w}}\sim\grave{\mathcal{Q}}}\left[\mathcal{L}_{\frac{\gamma}{2},\mathcal{E}}(f_{\widetilde{\mathbf{w}}})\right] \tag{1}$$

Also, for any $(h,r,t) \in \mathcal{E}$ and $\widetilde{\mathbf{w}} \sim \grave{\mathcal{Q}}$, we have

$$f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] - f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] \le \frac{\gamma}{2} \Rightarrow f_{\mathbf{w}}(h,r,t)[y_{hrt}] - f_{\mathbf{w}}(h,r,t)[1-y_{hrt}] \le \gamma$$
$$f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] \le \frac{\gamma}{2} + f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}] \Rightarrow f_{\mathbf{w}}(h,r,t)[y_{hrt}] \le \gamma + f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]$$

which indicates that

$$\mathbb{1}\big[f_{\widetilde{\mathbf{w}}}(h,r,t)[y_{hrt}] \le \frac{\gamma}{2} + f_{\widetilde{\mathbf{w}}}(h,r,t)[1-y_{hrt}]\big] \le \mathbb{1}\big[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \le \gamma + f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]\big]$$

This leads to $\mathcal{L}_{\frac{\gamma}{2},\widehat{\mathcal{E}}}(f_{\widetilde{\mathbf{w}}}) \le \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ for any $\widetilde{\mathbf{w}} \sim \grave{\mathcal{Q}}$. Then, we can end up with

$$\mathbb{E}_{\widetilde{\mathbf{w}}\sim\grave{\mathcal{Q}}}\big[\mathcal{L}_{\frac{\gamma}{2},\widehat{\mathcal{E}}}(f_{\widetilde{\mathbf{w}}})\big] \le \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) \tag{2}$$

Given $20 \le |\widehat{\mathcal{E}}| \le |\mathcal{E}| - 20$ and $|\mathcal{E}| \ge 40$, with probability $1-\delta$, we have

$$\begin{aligned}
\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \le& \mathbb{E}_{\widetilde{\mathbf{w}}\sim\grave{\mathcal{Q}}}\big[\mathcal{L}_{\frac{\gamma}{2},\mathcal{E}}(f_{\widetilde{\mathbf{w}}})\big] \quad \text{(Eq. (1))} \\
\le& \mathbb{E}_{\widetilde{\mathbf{w}}\sim\grave{\mathcal{Q}}}\big[\mathcal{L}_{\frac{\gamma}{2},\widehat{\mathcal{E}}}(f_{\widetilde{\mathbf{w}}})\big] + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[D_{KL}(\grave{\mathcal{Q}}\|\mathcal{P}) + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \quad \text{(Lemma C.1)} \\
\le& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[D_{KL}(\grave{\mathcal{Q}}\|\mathcal{P}) + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \quad \text{(Eq. (2))}
\end{aligned}$$

Also, we can derive the following.

$$\begin{aligned}
D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) =& \int_{\widetilde{\mathbf{w}}\in\mathcal{C}} \mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}} \ln\frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{\mathcal{P}} d\widetilde{\mathbf{w}} + \int_{\widetilde{\mathbf{w}}\in\mathcal{H}\setminus\mathcal{C}} \mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}} \ln\frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{\mathcal{P}} d\widetilde{\mathbf{w}} \\
=& p\int_{\widetilde{\mathbf{w}}\in\mathcal{C}} \frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{p} \ln\frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{p\mathcal{P}} d\widetilde{\mathbf{w}} + (1-p)\int_{\widetilde{\mathbf{w}}\in\mathcal{H}\setminus\mathcal{C}} \frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{1-p} \ln\frac{\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}}{(1-p)\mathcal{P}} d\widetilde{\mathbf{w}} \\
&+ \int_{\widetilde{\mathbf{w}}\in\mathcal{C}} \mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}} \ln p \, d\widetilde{\mathbf{w}} + \int_{\widetilde{\mathbf{w}}\in\mathcal{H}\setminus\mathcal{C}} \mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}} \ln(1-p) \, d\widetilde{\mathbf{w}} \\
=& pD_{KL}(\grave{\mathcal{Q}}\|\mathcal{P}) + (1-p)D_{KL}(\acute{\mathcal{Q}}\|\mathcal{P}) + p\ln p + (1-p)\ln(1-p)
\end{aligned}$$

Since we know $\frac{1}{2} < p < 1$ from the assumption, we have $(-\ln 2) < p\ln p + (1-p)\ln(1-p) < 0$. Also, D_{KL} is non-negative. Therefore, we have

$$\begin{aligned}
D_{KL}(\grave{\mathcal{Q}}\|\mathcal{P}) =& \frac{1}{p}(D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) - (1-p)D_{KL}(\acute{\mathcal{Q}}\|\mathcal{P}) - p\ln p - (1-p)\ln(1-p)) \\
\le& \frac{1}{p}(D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln 2) \le 2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + 2\ln 2
\end{aligned}$$

Finally, we show

$$\begin{aligned}
\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \le& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[D_{KL}(\grave{\mathcal{Q}}\|\mathcal{P}) + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \\
\le& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}
\end{aligned}$$

$\square$

D. Proofs of Theorem 4.4 and Theorem 4.5

We provide the complete proofs of Theorem 4.4 and Theorem 4.5.

Theorem 4.4 (Generalization Bound for ReED with Translational Distance Decoder)**.** *For any $L \geq 0$, let $f_{\mathbf{w}} : \mathcal{V} \times \mathcal{R} \times \mathcal{V} \to \mathbb{R}^2$ be a triplet classifier designed by the combination of the RAMP encoder with L-layers in Definition 3.1 and the TD decoder in Definition 3.2. Let k_r be the maximum of the infinity norms for all possible $\boldsymbol{S}_r^{(l)}$ in the RAMP encoder. Then, for any $\delta, \gamma > 0$, with probability at least $1-\delta$ over a training triplet set $\widehat{\mathcal{E}}$ (such that $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}| - 20$) sampled without replacement from the full triplet set $\mathcal{E}$, for any $\mathbf{w}$, we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}} L^2 \zeta_L^2 s^{2L} d \ln(N_{\mathbf{w}} d)}{\gamma^2} + \ln\frac{\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|)}{\delta}\right]}\right)$$

where $\theta(|\widehat{\mathcal{E}}|, |\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$, $N_{\mathbf{w}} = 2|\mathcal{R}|L + 6|\mathcal{R}| + 2L$, $\zeta_L = 2\tau^L\|\boldsymbol{X}_{\text{ent}}\|_2 + 2\kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{L-1}\tau^i\right) + \|\boldsymbol{X}_{\text{rel}}\|_2$, $\tau = C_\phi + \kappa, \kappa = C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} k_r, d = \max\left(\max_{0\leq l\leq L+1}(d_l), \max_{0\leq l\leq L+1}(d_l')\right), s_{L+1} = \max_{r,j}(\max(\|\overline{\boldsymbol{W}}_r^{\langle j\rangle}\|_F, \|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_F, \|\boldsymbol{V}_r^{\langle j\rangle}\|_F))$, $s_l = \max(\|\boldsymbol{W}_0^{(l)}\|_F, \|\boldsymbol{U}_0^{(l)}\|_F, \max_r\|\boldsymbol{W}_r^{(l)}\|_F, \max_r\|\boldsymbol{U}_r^{(l)}\|_F)$ for $l \in \{1, 2, \ldots, L\}$, and $s = \max_{1\leq l\leq L+1}(s_l)$.

Proof. We derive Theorem 4.4 from Theorem 4.3 where we construct a posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ by adding random perturbations $\ddot{\mathbf{w}}$ to $\mathbf{w}$. Following [24; 35], we set the prior distribution $\mathcal{P}$ as $\mathcal{N}\left(\mathbf{0}_{n_{\mathbf{w}}}, \sigma^2\boldsymbol{I}_{n_{\mathbf{w}}\times n_{\mathbf{w}}}\right)$ and the posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ as $\mathcal{N}\left(\mathbf{w}, \sigma^2\boldsymbol{I}_{n_{\mathbf{w}}\times n_{\mathbf{w}}}\right)$ where $n_{\mathbf{w}}$ is the size of $\mathbf{w}$. We first compute $\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty$, which we call the perturbation bound, so that we can calculate the standard deviation σ of the prior distribution that satisfies $\mathbb{P}\left(\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}\right) > \frac{1}{2}$. Afterwards, we calculate the KL divergence of $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ from $\mathcal{P}$ using σ and substitute the KL divergence term in Theorem 4.3 with our data and model-related terms, which finishes the proof.

Perturbation bound of ReED with translational distance decoder First, we compute the perturbation bound, $\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty$, and find σ that makes $\mathbb{P}(\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$ true. Let $\ddot{\boldsymbol{W}}$ denote a perturbation (also called noise) matrix added to the original weight matrix $\boldsymbol{W}$. As a result, we have $\widetilde{\boldsymbol{W}} = \boldsymbol{W} + \ddot{\boldsymbol{W}}$ where $\widetilde{\boldsymbol{W}}$ is a perturbed weight matrix. Also, let $\widetilde{\boldsymbol{H}}^{(l)}$ and $\widetilde{\boldsymbol{R}}^{(l)}$ denote the outputs of the perturbed model at the l-th layer. Each element of $\ddot{\boldsymbol{W}}$ is an i.i.d. element drawn from $\mathcal{N}\left(0, \sigma^2\right)$. Assume that the maximum of the Frobenius norms of the noise matrices is $\ddot{s}$. That is,

$$\ddot{s} = \max\left(\max_l\left(\left\|\ddot{\boldsymbol{W}}_0^{(l)}\right\|_F, \left\|\ddot{\boldsymbol{U}}_0^{(l)}\right\|_F, \max_r\left\|\ddot{\boldsymbol{W}}_r^{(l)}\right\|_F, \max_r\left\|\ddot{\boldsymbol{U}}_r^{(l)}\right\|_F\right), \quad \max_{r,j}\left(\left\|\ddot{\overline{\boldsymbol{W}}}_r^{\langle j\rangle}\right\|_F, \left\|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\right\|_F, \left\|\ddot{\boldsymbol{V}}_r^{\langle j\rangle}\right\|_F\right)\right)$$

Now, let us calculate the perturbation bound:

$$\begin{aligned}\max_{(h,r,t)\in\mathcal{E}}\|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty = \max_{(h,r,t)\in\mathcal{E}}\max_{j\in\{0,1\}}\Big| &- \|\widetilde{\boldsymbol{H}}^{(L)}[h,:]\widetilde{\overline{\boldsymbol{W}}}_r^{\langle j\rangle} + \widetilde{\boldsymbol{R}}^{(L)}[r,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} - \widetilde{\boldsymbol{H}}^{(L)}[t,:]\widetilde{\boldsymbol{V}}_r^{\langle j\rangle}\|_2 \\ &+ \|\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{W}}_r^{\langle j\rangle} + \boldsymbol{R}^{(L)}[r,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\boldsymbol{V}_r^{\langle j\rangle}\|_2\Big|\end{aligned}$$

Let us define $\Phi_l = \max_v \|\boldsymbol{H}^{(l)}[v,:]\|_2$, $\Psi_l = \max_v \|\widetilde{\boldsymbol{H}}^{(l)}[v,:] - \boldsymbol{H}^{(l)}[v,:]\|_2$, $\Lambda_l = \max_r \|\boldsymbol{R}^{(l)}[r,:]\|_2$, and $\Gamma_l = \max_r \|\widetilde{\boldsymbol{R}}^{(l)}[r,:] - \boldsymbol{R}^{(l)}[r,:]\|_2$. Then, for any $(h,r,t) \in \mathcal{E}$,

$$
\begin{aligned}
&\left| \left\| \widetilde{\boldsymbol{H}}^{(L)}[h,:]\widetilde{\overline{\boldsymbol{W}}}_r^{\langle j\rangle} + \widetilde{\boldsymbol{R}}^{(L)}[r,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} - \widetilde{\boldsymbol{H}}^{(L)}[t,:]\widetilde{\boldsymbol{V}}_r^{\langle j\rangle} \right\|_2 - \left\| \boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{W}}_r^{\langle j\rangle} + \boldsymbol{R}^{(L)}[r,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\boldsymbol{V}_r^{\langle j\rangle} \right\|_2 \right| \\
\leq& \left\| \widetilde{\boldsymbol{H}}^{(L)}[h,:]\widetilde{\overline{\boldsymbol{W}}}_r^{\langle j\rangle} + \widetilde{\boldsymbol{R}}^{(L)}[r,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} - \widetilde{\boldsymbol{H}}^{(L)}[t,:]\widetilde{\boldsymbol{V}}_r^{\langle j\rangle} \quad \text{(reverse triangle inequality)} \right. \\
&\left. - \left(\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{W}}_r^{\langle j\rangle} + \boldsymbol{R}^{(L)}[r,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\boldsymbol{V}_r^{\langle j\rangle} \right) \right\|_2 \\
=& \left\| \left(\widetilde{\boldsymbol{H}}^{(L)} - \boldsymbol{H}^{(L)} \right)[h,:]\widetilde{\overline{\boldsymbol{W}}}_r^{\langle j\rangle} + \left(\widetilde{\boldsymbol{R}}^{(L)} - \boldsymbol{R}^{(L)} \right)[r,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} \right. \\
&\left. - \left(\widetilde{\boldsymbol{H}}^{(L)} - \boldsymbol{H}^{(L)} \right)[t,:]\widetilde{\boldsymbol{V}}_r^{\langle j\rangle} + \boldsymbol{H}^{(L)}[h,:]\ddot{\overline{\boldsymbol{W}}}_r^{\langle 1\rangle} + \boldsymbol{R}^{(L)}[r,:]\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\ddot{\boldsymbol{V}}_r^{\langle j\rangle} \right\|_2 \\
\leq& \| \left(\widetilde{\boldsymbol{H}}^{(L)} - \boldsymbol{H}^{(L)} \right)[h,:]\|_2 \|\widetilde{\overline{\boldsymbol{W}}}_r^{\langle j\rangle}\|_2 + \| \left(\widetilde{\boldsymbol{R}}^{(L)} - \boldsymbol{R}^{(L)} \right)[r,:]\|_2 \|\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\|_2 \\
&+ \| \left(\widetilde{\boldsymbol{H}}^{(L)} - \boldsymbol{H}^{(L)} \right)[t,:]\|_2 \|\widetilde{\boldsymbol{V}}_r^{\langle j\rangle}\|_2 + \|\boldsymbol{H}^{(L)}[h,:]\|_2 \|\ddot{\overline{\boldsymbol{W}}}_r^{\langle j\rangle}\|_2 + \|\boldsymbol{R}^{(L)}[r,:]\|_2 \|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\|_2 \\
&+ \|\boldsymbol{H}^{(L)}[t,:]\|_2 \|\ddot{\boldsymbol{V}}_r^{\langle j\rangle}\|_2 \quad \text{(sub-additivity, sub-multiplicativity)} \\
\leq& (2\Psi_L + \Gamma_L)(s_{L+1} + \ddot{s}) + (2\Phi_L + \Lambda_L)\ddot{s} \quad \text{(definitions of } \Psi_L, \Gamma_L, s_{L+1}, \Phi_L, \Lambda_L, \text{and } \ddot{s}) \qquad (3)
\end{aligned}
$$

To satisfy the condition of Theorem 4.3, we need the bound of Φ_L, Ψ_L, Λ_L, and Γ_L. Since we need Λ_L and Γ_L to calculate Φ_L and Ψ_L, we first calculate the bound of Λ_L and Γ_L. In general, we need to compute ① bound of Λ_l and Γ_l, ② bound of Φ_l, and ③ bound of Ψ_l.

① Bound of Λ_l and Γ_l

First, we calculate the upper bound of $\|\boldsymbol{R}^{(l)}[r,:]\|_2$ and $\|\widetilde{\boldsymbol{R}}^{(l)}[r,:] - \boldsymbol{R}^{(l)}[r,:]\|_2$.

$$
\|\boldsymbol{R}^{(l)}[r,:]\|_2 = \|\boldsymbol{R}^{(0)}[r,:]\prod_{i=1}^{l}\boldsymbol{U}_0^{(i)}\|_2 \leq \|\boldsymbol{X}_{\text{rel}}\|_2 \prod_{i=1}^{l} s_i \qquad (4)
$$

$$
\begin{aligned}
\|\widetilde{\boldsymbol{R}}^{(l)}[r,:] - \boldsymbol{R}^{(l)}[r,:]\|_2 =& \|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\widetilde{\boldsymbol{U}}_0^{(l)} - \boldsymbol{R}^{(l-1)}[r,:]\boldsymbol{U}_0^{(l)}\|_2 \\
=& \|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\widetilde{\boldsymbol{U}}_0^{(l)} - \boldsymbol{R}^{(l-1)}[r,:]\widetilde{\boldsymbol{U}}_0^{(l)} + \boldsymbol{R}^{(l-1)}[r,:]\ddot{\boldsymbol{U}}_0^{(l)}\|_2 \\
\leq& \|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:] - \boldsymbol{R}^{(l-1)}[r,:]\|_2 \|\widetilde{\boldsymbol{U}}_0^{(l)}\|_2 \\
&+ \|\boldsymbol{R}^{(l-1)}[r,:]\|_2 \|\ddot{\boldsymbol{U}}_0^{(l)}\|_2 \quad \text{(sub-additivity, sub-multiplicativity)} \\
\leq& \|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:] - \boldsymbol{R}^{(l-1)}[r,:]\|_2 (s_l + \ddot{s}) + \|\boldsymbol{R}^{(l-1)}[r,:]\|_2 \ddot{s} \quad \text{(definitions of } s_l \text{ and } \ddot{s}) \\
\leq& \|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:] - \boldsymbol{R}^{(l-1)}[r,:]\|_2 (s_l + \ddot{s}) + \ddot{s}\|\boldsymbol{X}_{\text{rel}}\|_2 \prod_{i=1}^{l-1} s_i \quad \text{(Eq.(4))}
\end{aligned}
$$

Note that

$$
\begin{aligned}
\|\widetilde{\boldsymbol{R}}^{(1)}[r,:] - \boldsymbol{R}^{(1)}[r,:]\|_2 =& \|\boldsymbol{X}_{\text{rel}}[r,:]\widetilde{\boldsymbol{U}}_0^{(1)} - \boldsymbol{X}_{\text{rel}}[r,:]\boldsymbol{U}_0^{(1)}\|_2 \\
=& \|\boldsymbol{X}_{\text{rel}}[r,:]\ddot{\boldsymbol{U}}_0^{(1)}\|_2 \\
\leq& \|\boldsymbol{X}_{\text{rel}}[r,:]\|_2 \|\ddot{\boldsymbol{U}}_0^{(1)}\|_2 \\
\leq& \|\boldsymbol{X}_{\text{rel}}\|_2 \ddot{s} \quad \text{(definitions of the spectral norm and } \ddot{s}) \qquad (5)
\end{aligned}
$$

Then, we get

$$\begin{aligned}
\|\widetilde{\boldsymbol{R}}^{(l)}[r,:]-\boldsymbol{R}^{(l)}[r,:]\|_2+\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l}s_i \leq& \left(\|\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]-\boldsymbol{R}^{(l-1)}[r,:]\|_2+\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l-1}s_i\right)(s_l+\ddot{s})\\
\leq&\Big(\|\widetilde{\boldsymbol{R}}^{(1)}[r,:]-\boldsymbol{R}^{(1)}[r,:]\|_2+\|\boldsymbol{X}_{\text{rel}}\|_2 s_1\Big)\prod_{i=2}^{l}(s_i+\ddot{s}) \quad \text{(recursion)}\\
\leq&\left(\|\boldsymbol{X}_{\text{rel}}\|_2\ddot{s}+\|\boldsymbol{X}_{\text{rel}}\|_2 s_1\right)\prod_{i=2}^{l}(s_i+\ddot{s}) \quad \text{(Eq. (5))}\\
=&\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l}(s_i+\ddot{s})
\end{aligned}$$

Putting all this together, we have

$$\Lambda_l=\max_r\|\boldsymbol{R}^{(l)}[r,:]\|_2\leq\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l}s_i \tag{6}$$

$$\Gamma_l=\max_r\|\widetilde{\boldsymbol{R}}^{(l)}[r,:]-\boldsymbol{R}^{(l)}[r,:]\|_2\leq\|\boldsymbol{X}_{\text{rel}}\|_2\left(\prod_{i=1}^{l}(s_i+\ddot{s})-\prod_{i=1}^{l}s_i\right) \tag{7}$$

Note that the bound of Λ_L and Γ_L can be obtained by setting $l=L$.

② Bound of Φ_l

We can calculate the upper bound of Φ_l using sub-additivity, sub-multiplicativity, and the definitions of matrix norms. Note that the bound of Φ_L can be obtained by setting $l=L$.

Let $v^*=\operatorname{argmax}_v\|\boldsymbol{H}^{(l)}[v,:]\|_2$. Then,

$$\begin{aligned}
\Phi_l=&\|\phi\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}+\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\right)[v^*,:]\|_2\\
\leq&C_\phi\|\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^*,:]+\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^*,:]\|_2 \quad \text{(Lipschitzness of }\phi,\ \phi(\mathbf{0})=\mathbf{0})\\
\leq&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2+C_\phi\|\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^*,:]\|_2 \quad \text{(sub-additivity, sub-multiplicativity)}\\
\leq&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2+C_\phi C_\rho\|\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^*,:]\|_2 \quad \text{(Lipschitzness of }\rho,\ \rho(\mathbf{0})=\mathbf{0})\\
=&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2+C_\phi C_\rho\|\sum_{r\in\mathcal{R}}\left(\sum_{v\in\mathcal{V}}\boldsymbol{S}_r^{(l)}[v^*,v]\left(\psi\left(\boldsymbol{M}_r^{(l)}\right)[v,:]\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\|_2\\
=&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2 \qquad \text{(definition of }\boldsymbol{M}_r^{(l)})\\
&+C_\phi C_\rho\|\sum_{r\in\mathcal{R}}\left(\sum_{v\in\mathcal{V}}\boldsymbol{S}_r^{(l)}[v^*,v]\left(\psi\left(\begin{bmatrix}\boldsymbol{H}^{(l-1)}[v,:] & \boldsymbol{R}^{(l-1)}[r,:]\end{bmatrix}\right)\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\|_2\\
\leq&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2 \qquad \text{(sub-additivity, absolute homogeneity)}\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^*,v]\right|\|\left(\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\boldsymbol{W}_r^{(l)}+\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\boldsymbol{U}_r^{(l)}\right)\|_2\\
\leq&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2 \qquad \text{(sub-additivity, sub-multiplicativity)}\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^*,v]\right|\left(\|\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\|_2\|\boldsymbol{W}_r^{(l)}\|_2+\|\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\|_2\|\boldsymbol{U}_r^{(l)}\|_2\right)\\
\leq&C_\phi\|\boldsymbol{H}^{(l-1)}[v^*,:]\|_2\|\boldsymbol{W}_0^{(l)}\|_2 \quad \text{(Lipschitzness of }\psi,\ \psi(\mathbf{0})=\mathbf{0})
\end{aligned}$$

$$
\begin{aligned}
&+ C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}} \left|\boldsymbol{S}_r^{(l)}[v^*,v]\right| \left(\|\boldsymbol{H}^{(l-1)}[v,:]\|_2\|\boldsymbol{W}_r^{(l)}\|_2 + \|\boldsymbol{R}^{(l-1)}[r,:]\|_2\|\boldsymbol{U}_r^{(l)}\|_2\right) \\
\le& C_\phi \Phi_{l-1}\|\boldsymbol{W}_0^{(l)}\|_2 + C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} \left(\left(\Phi_{l-1}\|\boldsymbol{W}_r^{(l)}\|_2 + \Lambda_{l-1}\|\boldsymbol{U}_r^{(l)}\|_2\right)\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^*,v]\right|\right) \quad \text{(definitions of } \Phi_l \text{ and } \Lambda_l\text{)} \\
\le& C_\phi \Phi_{l-1}\|\boldsymbol{W}_0^{(l)}\|_2 + C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} \left(\Phi_{l-1}\|\boldsymbol{W}_r^{(l)}\|_2 + \Lambda_{l-1}\|\boldsymbol{U}_r^{(l)}\|_2\right)\|\boldsymbol{S}_r^{(l)}\|_\infty \quad \text{(definition of the infinity norm of a matrix)} \\
\le& C_\phi \Phi_{l-1}\|\boldsymbol{W}_0^{(l)}\|_2 + C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} \left(\Phi_{l-1}\|\boldsymbol{W}_r^{(l)}\|_2 + \Lambda_{l-1}\|\boldsymbol{U}_r^{(l)}\|_2\right) k_r \quad \text{(definition of } k_r\text{)} \\
\le& C_\phi s_l \Phi_{l-1} + C_\phi C_\rho C_\psi s_l(\Phi_{l-1}+\Lambda_{l-1})\sum_{r\in\mathcal{R}} k_r \quad \text{(definition of } s_l\text{)} \\
\le& C_\phi s_l \Phi_{l-1} + C_\phi C_\rho C_\psi s_l\left(\Phi_{l-1} + \|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l-1}s_i\right)\sum_{r\in\mathcal{R}} k_r \quad \text{(Eq.(6))} \\
\le& \tau s_l\Phi_{l-1} + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l}s_i \quad \text{(definitions of } \tau \text{ and } \kappa\text{)}
\end{aligned} \tag{8}
$$

Note that $\Phi_1 \le \tau s_1\Phi_0 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2 s_1 \le \tau s_1\|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2 s_1$ by Eq. (8) and the definition of the spectral norm. Then,

$$
\left(\Phi_l - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\prod_{i=1}^{l}s_i\right) - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=1}^{l-1}\tau^i\right)\left(\prod_{i=1}^{l}s_i\right) \le \tau s_l\Phi_{l-1} - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=1}^{l-1}\tau^i\right)\left(\prod_{i=1}^{l}s_i\right)
$$

which leads to

$$
\begin{aligned}
\Phi_l - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{l-1}\tau^i\right)\left(\prod_{i=1}^{l}s_i\right) \le& \tau s_l\left(\Phi_{l-1} - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{l-2}\tau^i\right)\left(\prod_{i=1}^{l-1}s_i\right)\right) \\
\le& \tau^{l-1}\left(\prod_{i=2}^{l}s_i\right)\left(\Phi_1 - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{0}\tau^i\right)\left(\prod_{i=1}^{1}s_i\right)\right) \quad \text{(recursion)} \\
\le& \tau^{l-1}\left(\prod_{i=2}^{l}s_i\right)\left(\tau s_1\|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2 s_1 - \kappa\|\boldsymbol{X}_{\text{rel}}\|_2 s_1\right) \\
\le& \tau^l\|\boldsymbol{X}_{\text{ent}}\|_2\prod_{i=1}^{l}s_i
\end{aligned}
$$

Finally, we get

$$
\Phi_l \le \tau^l\|\boldsymbol{X}_{\text{ent}}\|_2\prod_{i=1}^{l}s_i + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\left(\sum_{i=0}^{l-1}\tau^i\right)\left(\prod_{i=1}^{l}s_i\right) = \left(\tau^l\|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\sum_{i=0}^{l-1}\tau^i\right)\prod_{i=1}^{l}s_i = \eta_l\prod_{i=1}^{l}s_i \tag{9}
$$

where $\eta_l = \tau^l\|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\text{rel}}\|_2\sum_{i=0}^{l-1}\tau^i$.

③ Bound of Ψ_l

We can calculate the upper bound of Ψ_l using sub-additivity, sub-multiplicativity, and the definitions of matrix norms. Note that the bound of Ψ_L can be obtained by setting $l = L$.

Let $v^{**} = \text{argmax}_v \|\widetilde{\boldsymbol{H}}^{(l)}[v,:] - \boldsymbol{H}^{(l)}[v,:]\|_2$. Then,

$$
\Psi_l = \left\|\phi\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)} + \rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right)\right)[v^{**},:]\right.
$$

$$
\begin{aligned}
&-\phi\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}+\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\right)[v^{**},:]\Bigg\|_2\\
\le& C_\phi\left\|\left(\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]+\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right)[v^{**},:]\right)\right. \quad \text{(Lipschitzness of } \phi)\\
&\left.-\left(\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]+\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^{**},:]\right)\right\|_2\\
\le& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2 \quad \text{(sub-additivity)}\\
&+C_\phi\|\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right)[v^{**},:]-\rho\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^{**},:]\|_2\\
\le& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2 \quad \text{(Lipschitzness of } \rho)\\
&+C_\phi C_\rho\|\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right)[v^{**},:]-\left(\sum_{r\in\mathcal{R}}\boldsymbol{S}_r^{(l)}\psi\left(\boldsymbol{M}_r^{(l)}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)[v^{**},:]\|_2\\
=& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2\\
&+C_\phi C_\rho\left\|\left(\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\boldsymbol{S}_r^{(l)}[v^{**},v]\left(\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)[v,:]\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right)-\left(\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\boldsymbol{S}_r^{(l)}[v^{**},v]\left(\psi\left(\boldsymbol{M}_r^{(l)}\right)[v,:]\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\right\|_2\\
=& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2\\
&+C_\phi C_\rho\|\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\boldsymbol{S}_r^{(l)}[v^{**},v]\left(\left(\psi\left(\widetilde{\boldsymbol{M}}_r^{(l)}\right)[v,:]\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}-\left(\psi\left(\boldsymbol{M}_r^{(l)}\right)[v,:]\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right)\|_2\\
\le& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2 \quad \text{(definition of } \boldsymbol{M}_r^{(l)} \text{ and sub-additivity)}\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^{**},v]\right|\left\|\psi\left(\begin{bmatrix}\widetilde{\boldsymbol{H}}^{(l-1)}[v,:] & \widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\end{bmatrix}\right)\begin{bmatrix}\widetilde{\boldsymbol{W}}_r^{(l)}\\ \widetilde{\boldsymbol{U}}_r^{(l)}\end{bmatrix}\right.\\
&\qquad\left.-\psi\left(\begin{bmatrix}\boldsymbol{H}^{(l-1)}[v,:] & \boldsymbol{R}^{(l-1)}[r,:]\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{W}_r^{(l)}\\ \boldsymbol{U}_r^{(l)}\end{bmatrix}\right\|_2 \quad \text{(absolute homogeneity)}\\
=& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]-\left(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_0^{(l)}\right)[v^{**},:]\|_2\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^{**},v]\right|\left\|\psi\left(\widetilde{\boldsymbol{H}}^{(l-1)}[v,:]\right)\widetilde{\boldsymbol{W}}_r^{(l)}+\psi\left(\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\right)\widetilde{\boldsymbol{U}}_r^{(l)}\right.\\
&\qquad\left.-\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\boldsymbol{W}_r^{(l)}-\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\boldsymbol{U}_r^{(l)}\right\|_2\\
=& C_\phi\|\left(\left(\widetilde{\boldsymbol{H}}^{(l-1)}-\boldsymbol{H}^{(l-1)}\right)\widetilde{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]+\left(\boldsymbol{H}^{(l-1)}\ddot{\boldsymbol{W}}_0^{(l)}\right)[v^{**},:]\|_2 \quad \text{(definition of } \widetilde{\boldsymbol{W}})\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^{**},v]\right|\left\|\left(\psi\left(\widetilde{\boldsymbol{H}}^{(l-1)}[v,:]\right)-\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\right)\widetilde{\boldsymbol{W}}_r^{(l)}\right.\\
&\qquad+\left(\psi\left(\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\right)-\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\right)\widetilde{\boldsymbol{U}}_r^{(l)}\\
&\qquad\left.+\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\ddot{\boldsymbol{W}}_r^{(l)}+\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\ddot{\boldsymbol{U}}_r^{(l)}\right\|_2\\
\le& C_\phi\|\left(\widetilde{\boldsymbol{H}}^{(l-1)}-\boldsymbol{H}^{(l-1)}\right)[v^{**},:]\|_2\|\widetilde{\boldsymbol{W}}_0^{(l)}\|_2+C_\phi\|\boldsymbol{H}^{(l-1)}[v^{**},:]\|_2\|\ddot{\boldsymbol{W}}_0^{(l)}\|_2 \quad \text{(sub-additivity, sub-multiplicativity)}\\
&+C_\phi C_\rho\sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{V}}\left|\boldsymbol{S}_r^{(l)}[v^{**},v]\right|\left(\|\psi\left(\widetilde{\boldsymbol{H}}^{(l-1)}[v,:]\right)-\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\|_2\|\widetilde{\boldsymbol{W}}_r^{(l)}\|_2\right.\\
&\qquad+\|\psi\left(\widetilde{\boldsymbol{R}}^{(l-1)}[r,:]\right)-\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\|_2\|\widetilde{\boldsymbol{U}}_r^{(l)}\|_2\\
&\qquad\left.+\|\psi\left(\boldsymbol{H}^{(l-1)}[v,:]\right)\|_2\|\ddot{\boldsymbol{W}}_r^{(l)}\|_2+\|\psi\left(\boldsymbol{R}^{(l-1)}[r,:]\right)\|_2\|\ddot{\boldsymbol{U}}_r^{(l)}\|_2\right)
\end{aligned}
$$

$$
\begin{aligned}
\leq & C_\phi \| \left(\widetilde{\boldsymbol{H}}^{(l-1)} - \boldsymbol{H}^{(l-1)} \right) [v^{**}, :] \|_2 \| \widetilde{\boldsymbol{W}}_0^{(l)} \|_2 + C_\phi \| \boldsymbol{H}^{(l-1)} [v^{**}, :] \|_2 \| \ddot{\boldsymbol{W}}_0^{(l)} \|_2 \qquad \text{(Lipschitzness of } \psi, \psi(\mathbf{0}) = \mathbf{0}) \\
& + C_\phi C_\rho C_\psi \sum_{r \in \mathcal{R}} \sum_{v \in \mathcal{V}} \left| \boldsymbol{S}_r^{(l)} [v^{**}, v] \right| \Big(\| \widetilde{\boldsymbol{H}}^{(l-1)} [v, :] - \boldsymbol{H}^{(l-1)} [v, :] \|_2 \| \widetilde{\boldsymbol{W}}_r^{(l)} \|_2 \\
& \qquad + \| \widetilde{\boldsymbol{R}}^{(l-1)} [r, :] - \boldsymbol{R}^{(l-1)} [r, :] \|_2 \| \widetilde{\boldsymbol{U}}_r^{(l)} \|_2 \\
& \qquad + \| \boldsymbol{H}^{(l-1)} [v, :] \|_2 \| \ddot{\boldsymbol{W}}_r^{(l)} \|_2 + \| \boldsymbol{R}^{(l-1)} [r, :] \|_2 \| \ddot{\boldsymbol{U}}_r^{(l)} \|_2 \Big) \\
\leq & C_\phi (s_l + \ddot{s}) \Psi_{l-1} + C_\phi \Phi_{l-1} \ddot{s} \qquad \text{(definitions of } \Phi_l, \Psi_l, \Lambda_l, \Gamma_l, s_l, \text{ and } \ddot{s}) \\
& + C_\phi C_\rho C_\psi \sum_{r \in \mathcal{R}} \sum_{v \in \mathcal{V}} \left| \boldsymbol{S}_r^{(l)} [v^{**}, v] \right| \left(\Psi_{l-1} (s_l + \ddot{s}) + \Gamma_{l-1} (s_l + \ddot{s}) + \Phi_{l-1} \ddot{s} + \Lambda_{l-1} \ddot{s} \right) \\
\leq & C_\phi (s_l + \ddot{s}) \Psi_{l-1} + C_\phi \Phi_{l-1} \ddot{s} \qquad \text{(definition of the infinity norm of a matrix)} \\
& + C_\phi C_\rho C_\psi \left(\Psi_{l-1} (s_l + \ddot{s}) + \Gamma_{l-1} (s_l + \ddot{s}) + \Phi_{l-1} \ddot{s} + \Lambda_{l-1} \ddot{s} \right) \sum_{r \in \mathcal{R}} \| \boldsymbol{S}_r^{(l)} \|_\infty \\
= & C_\phi (s_l + \ddot{s}) \Psi_{l-1} + C_\phi \Phi_{l-1} \ddot{s} \\
& + C_\phi C_\rho C_\psi (\Psi_{l-1} + \Gamma_{l-1}) (s_l + \ddot{s}) \sum_{r \in \mathcal{R}} \| \boldsymbol{S}_r \|_\infty + C_\phi C_\rho C_\psi (\Phi_{l-1} + \Lambda_{l-1}) \ddot{s} \sum_{r \in \mathcal{R}} \| \boldsymbol{S}_r^{(l)} \|_\infty \\
\leq & C_\phi (s_l + \ddot{s}) \Psi_{l-1} + C_\phi \Phi_{l-1} \ddot{s} \qquad \text{(definition of } k_r) \\
& + C_\phi C_\rho C_\psi (\Psi_{l-1} + \Gamma_{l-1}) (s_l + \ddot{s}) \sum_{r \in \mathcal{R}} k_r + C_\phi C_\rho C_\psi (\Phi_{l-1} + \Lambda_{l-1}) \ddot{s} \sum_{r \in \mathcal{R}} k_r \\
\leq & \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau \Phi_{l-1} \ddot{s} + \kappa (s_l + \ddot{s}) \Gamma_{l-1} + \kappa \ddot{s} \Lambda_{l-1} \qquad \text{(definitions of } \tau \text{ and } \kappa) \qquad (10) \\
\leq & \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau \ddot{s} \eta_{l-1} \prod_{i=1}^{l-1} s_i + \kappa (s_l + \ddot{s}) \| \boldsymbol{X}_{\text{rel}} \|_2 \left(\prod_{i=1}^{l-1} (s_i + \ddot{s}) - \prod_{i=1}^{l-1} s_i \right) + \kappa \ddot{s} \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l-1} s_i \qquad \text{(Eq. (6), (7), (9))} \\
= & \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau \ddot{s} \eta_{l-1} \prod_{i=1}^{l-1} s_i + \kappa (s_l + \ddot{s}) \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l-1} (s_i + \ddot{s}) - \kappa s_l \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l-1} s_i \\
= & \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau \ddot{s} \eta_{l-1} \prod_{i=1}^{l-1} s_i + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} (s_i + \ddot{s}) - \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} s_i \\
= & \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau \ddot{s} \eta_{l-1} \prod_{i=1}^{l-1} s_i + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \left(\prod_{i=1}^{l} (s_i + \ddot{s}) - \prod_{i=1}^{l} s_i \right)
\end{aligned}
$$

From the above inequality, we can induce

$$
\Psi_l + \tau \eta_{l-1} \prod_{i=1}^{l} s_i + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} s_i \leq \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau (s_l + \ddot{s}) \eta_{l-1} \prod_{i=1}^{l-1} s_i + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} (s_i + \ddot{s})
$$

where

$$
\Psi_l + \tau \eta_{l-1} \prod_{i=1}^{l} s_i + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} s_i = \Psi_l + (\tau \eta_{l-1} + \kappa \| \boldsymbol{X}_{\text{rel}} \|_2) \prod_{i=1}^{l} s_i = \Psi_l + \eta_l \prod_{i=1}^{l} s_i
$$

Putting all this together, we have

$$
\begin{aligned}
& \Psi_l + \eta_l \prod_{i=1}^{l} s_i - \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \prod_{i=1}^{l} (s_i + \ddot{s}) - \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \left(\sum_{i=1}^{l-1} \tau^i \right) \left(\prod_{i=1}^{l} (s_i + \ddot{s}) \right) \\
& \leq \tau (s_l + \ddot{s}) \Psi_{l-1} + \tau (s_l + \ddot{s}) \eta_{l-1} \prod_{i=1}^{l-1} s_i - \kappa \| \boldsymbol{X}_{\text{rel}} \|_2 \left(\sum_{i=1}^{l-1} \tau^i \right) \left(\prod_{i=1}^{l} (s_i + \ddot{s}) \right)
\end{aligned}
$$

Note that $\Psi_1 \leq \tau (s_1 + \ddot{s}) \Psi_0 + \tau \ddot{s} \Phi_0 + \kappa (s_1 + \ddot{s}) \Gamma_0 + \kappa \ddot{s} \Lambda_0 \leq \tau \ddot{s} \| \boldsymbol{X}_{\text{ent}} \|_2 + \kappa \ddot{s} \| \boldsymbol{X}_{\text{rel}} \|_2$ by Eq. (10) and the definitions of $\Psi_0, \Phi_0, \Lambda_0$ and Γ_0.

Gathering up,

$$
\begin{aligned}
&\Psi_l + \eta_l \prod_{i=1}^{l} s_i - \kappa \|\boldsymbol{X}_{\text{rel}}\|_2 \left(\sum_{i=0}^{l-1} \tau^i\right) \left(\prod_{i=1}^{l} (s_i + \ddot{s})\right) \\
&\leq \tau(s_l + \ddot{s}) \left(\Psi_{l-1} + \eta_{l-1} \prod_{i=1}^{l-1} s_i - \kappa \|\boldsymbol{X}_{\text{rel}}\|_2 \left(\sum_{i=0}^{l-2} \tau^i\right) \left(\prod_{i=1}^{l-1} (s_i + \ddot{s})\right)\right) \\
&\leq \tau^{l-1} \left(\prod_{i=2}^{l} (s_i + \ddot{s})\right) \left(\Psi_1 + \eta_1 \prod_{i=1}^{1} s_i - \kappa \|\boldsymbol{X}_{\text{rel}}\|_2 \left(\sum_{i=0}^{0} \tau^i\right) \left(\prod_{i=1}^{1} (s_i + \ddot{s})\right)\right) \qquad \text{(recursion)} \\
&\leq \tau^{l-1} \left(\prod_{i=2}^{l} (s_i + \ddot{s})\right) (\ddot{s}(\tau \|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa \|\boldsymbol{X}_{\text{rel}}\|_2) + (\tau \|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa \|\boldsymbol{X}_{\text{rel}}\|_2) s_1 - \kappa \|\boldsymbol{X}_{\text{rel}}\|_2 (s_1 + \ddot{s})) \\
&= \tau^l \left(\prod_{i=1}^{l} (s_i + \ddot{s})\right) \|\boldsymbol{X}_{\text{ent}}\|_2
\end{aligned}
$$

We end up with

$$
\begin{aligned}
\Psi_l \leq& \tau^l \left(\prod_{i=1}^{l} (s_i + \ddot{s})\right) \|\boldsymbol{X}_{\text{ent}}\|_2 + \kappa \|\boldsymbol{X}_{\text{rel}}\|_2 \left(\sum_{i=0}^{l-1} \tau^i\right) \left(\prod_{i=1}^{l} (s_i + \ddot{s})\right) - \eta_l \prod_{i=1}^{l} s_i \\
=& \eta_l \left(\prod_{i=1}^{l} (s_i + \ddot{s}) - \prod_{i=1}^{l} s_i\right) \qquad \text{(definition of } \eta_l)
\end{aligned} \tag{11}
$$

Then,

$$
\begin{aligned}
\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty \leq& (2\Psi_L + \Gamma_L)(s_{L+1} + \ddot{s}) + (2\Phi_L + \Lambda_L)\ddot{s} \qquad \text{(Eq. (3))} \\
\leq& \left(\prod_{i=1}^{L} (s_i + \ddot{s}) - \prod_{i=1}^{L} s_i\right) (2\eta_L + \|\boldsymbol{X}_{\text{rel}}\|_2)(s_{L+1} + \ddot{s}) \quad \text{(Eq. (6), (7), (9), (11))} \\
&+ \left(\prod_{i=1}^{L} s_i\right) (2\eta_L + \|\boldsymbol{X}_{\text{rel}}\|_2)\ddot{s} \\
=& \left(\prod_{i=1}^{L+1} (s_i + \ddot{s}) - \prod_{i=1}^{L+1} s_i\right) (2\eta_L + \|\boldsymbol{X}_{\text{rel}}\|_2) \\
\leq& \left((s + \ddot{s})^{L+1} - s^{L+1}\right) \zeta_L \qquad \text{(definition of } s) \\
\leq& \ddot{s}(L+1)(s + \ddot{s})^L \zeta_L \qquad (0 \leq s \leq s + \ddot{s})
\end{aligned} \tag{12}
$$

where $\zeta_l = 2\eta_l + \|\boldsymbol{X}_{\text{rel}}\|_2$.

Generalization bound of ReED with translational distance decoder Recall that we set the prior distribution $\mathcal{P}$ as $\mathcal{N}\left(\mathbf{0}_{n_\mathbf{w}}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w} \times n_\mathbf{w}}\right)$. Since the distribution of the perturbed parameters $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ is $\mathcal{N}\left(\mathbf{w}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w} \times n_\mathbf{w}}\right)$, the distribution of the perturbation is $\mathcal{N}\left(\mathbf{0}_{n_\mathbf{w}}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w} \times n_\mathbf{w}}\right)$. For the perturbation matrices that follow the normal distribution, we can derive the following by replacing the standard normal variable in Corollary 4.2. in Tropp [48] with a random variable following the normal distribution having the mean of 0 and the variance of σ^2.

$$\begin{aligned}
&\mathbb{P}\left(\left\|\ddot{\boldsymbol{W}}_0^{(l)}\right\|_2 \geq \ddot{s}\right) \leq (d_{l-1}+d_l)\, e^{-\ddot{s}^2/(2\max(d_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\boldsymbol{U}}_0^{(l)}\right\|_2 \geq \ddot{s}\right) \leq (d'_{l-1}+d'_l)e^{-\ddot{s}^2/(2\max(d'_{l-1},d'_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\boldsymbol{W}}_r^{(l)}\right\|_2 \geq \ddot{s}\right) \leq (d_{l-1}+d_l)e^{-\ddot{s}^2/(2\max(d_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\boldsymbol{U}}_r^{(l)}\right\|_2 \geq \ddot{s}\right) \leq (d'_{l-1}+d_l)e^{-\ddot{s}^2/(2\max(d'_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\overline{\boldsymbol{W}}}_r^{\langle j\rangle}\right\|_2 \geq \ddot{s}\right) \leq (d_L+d_{L+1})e^{-\ddot{s}^2/(2\max(d_L,d_{L+1})\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\right\|_2 \geq \ddot{s}\right) \leq (d'_L+d_{L+1})e^{-\ddot{s}^2/(2\max(d'_L,d_{L+1})\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}\\
&\mathbb{P}\left(\left\|\ddot{\boldsymbol{V}}_r^{\langle j\rangle}\right\|_2 \geq \ddot{s}\right) \leq (d_L+d_{L+1})e^{-\ddot{s}^2/(2\max(d_L,d_{L+1})\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}
\end{aligned}$$

Using Bernoulli's inequality, we can derive that the probability of all perturbation matrices having the spectral norm less than $\ddot{s}$ is greater than or equal to $1-2N_{\mathbf{w}}de^{-\ddot{s}^2/(2d\sigma^2)}$, where $N_{\mathbf{w}} = 2\cdot L + 2\cdot|\mathcal{R}|L + 3\cdot 2|\mathcal{R}| = 2|\mathcal{R}|L + 6|\mathcal{R}| + 2L$ is the number of perturbation matrices.

To satisfy the condition of Theorem 4.3, we set $2N_{\mathbf{w}}de^{-\ddot{s}^2/(2d\sigma^2)} = 1/2$. Then, we get $\ddot{s} = \sigma\sqrt{2d\ln(4N_{\mathbf{w}}d)}$. Since the prior is independent of the learned parameters $\mathbf{w}$, we cannot directly use s to formulate σ. Therefore, we approximate s with $\mathring{s}$ in the following range.

$$|s-\mathring{s}| \leq \frac{1}{L+2}s \implies \frac{L+1}{L+2}s \leq \mathring{s} \leq \frac{L+3}{L+2}s$$

Additionally, we assume that $\ddot{s} \leq \frac{1}{L+2}s$. Then, if

$$\begin{aligned}
\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty \leq& \ddot{s}(L+1)(s+\ddot{s})^L\zeta_L \qquad \text{(Eq. (12))}\\
\leq& \ddot{s}(L+1)s^L\left(1+\frac{1}{L+2}\right)^L\zeta_L \qquad \left(\ddot{s}\leq\frac{1}{L+2}s\right)\\
\leq& \ddot{s}(L+1)\left(\frac{L+2}{L+1}\right)^L\mathring{s}^L\left(1+\frac{1}{L+2}\right)^L\zeta_L \ \text{(range of } \mathring{s}\text{)}\\
=& \ddot{s}(L+1)\mathring{s}^L\left(1+\frac{2}{L+1}\right)^L\zeta_L\\
\leq& \ddot{s}(L+1)\mathring{s}^Le^2\zeta_L \leq \frac{\gamma}{4} \qquad \left(\left(1+\frac{1}{x}\right)^x \leq e, \forall x\geq 0\right)
\end{aligned}$$

is satisfied, we meet the condition of Theorem 4.3. With $\ddot{s} = \sigma\sqrt{2d\ln(4N_{\mathbf{w}}d)}$, we have

$$\begin{aligned}
\ddot{s} = \sigma\sqrt{2d\ln(4N_{\mathbf{w}}d)} \leq& \frac{\gamma}{4e^2(L+1)\mathring{s}^L\zeta_L}\\
\rightarrow \sigma \leq& \frac{1}{\sqrt{2d\ln(4N_{\mathbf{w}}d)}}\left(\frac{\gamma}{4e^2(L+1)\mathring{s}^L\zeta_L}\right)
\end{aligned}$$

By setting $\sigma = \frac{1}{\sqrt{2d\ln(4N_{\mathbf{w}}d)}}\left(\frac{\gamma}{4e^2(L+1)\mathring{s}^L\zeta_L}\right)$, we can calculate $D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P})$.

$$\begin{aligned}
D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) &= \frac{\|\mathbf{w}\|_2^2}{2\sigma^2} \qquad \text{(KL divergence between two normal distributions)}\\
&= \frac{2d\ln(4N_{\mathbf{w}}d)}{2\left(\frac{\gamma}{4e^2(L+1)\mathring{s}^L\zeta_L}\right)^2}\left(\sum\|\boldsymbol{W}\|_F^2\right)\\
&\leq \frac{\left(4e^2(L+1)\mathring{s}^L\zeta_L\right)^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2}\left(N_{\mathbf{w}}s^2\right)\\
&= \frac{16e^4N_{\mathbf{w}}(L+1)^2\mathring{s}^{2L}s^2\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2}\\
&\leq \frac{16e^4N_{\mathbf{w}}(L+1)^2 s^2\left(\frac{L+3}{L+2}s\right)^{2L}\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} \qquad \text{(range of } \mathring{s}\text{)}\\
&\leq \frac{16e^6N_{\mathbf{w}}(L+1)^2 s^{2L+2}\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} \qquad \left(\left(1+\frac{1}{x}\right)^x \leq e, \forall x \geq 0\right)
\end{aligned}$$

From Theorem 4.3, we get

$$\begin{aligned}
\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\\
\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6N_{\mathbf{w}}(L+1)^2 s^{2L+2}\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} + \frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \qquad (13)
\end{aligned}$$

Now, let us find some range of s such that Theorem 4.4 trivially holds.

First, if

$$\begin{aligned}
\|f_{\mathbf{w}}(h,r,t)\|_\infty &= \max_{(h,r,t)\in\mathcal{E}}\max_{j\in\{0,1\}}\left|-\|\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{W}}_r^{\langle j\rangle} + \boldsymbol{R}^{(L)}[r,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} - \boldsymbol{H}^{(L)}[t,:]\boldsymbol{V}_r^{\langle j\rangle}\|_2\right|\\
&\leq \max_{(h,r,t)\in\mathcal{E}}\max_{j\in\{0,1\}}\Big(\|\boldsymbol{H}^{(L)}[h,:]\|_2\|\overline{\boldsymbol{W}}_r^{\langle j\rangle}\|_2 \qquad \text{(sub-additivity, sub-multiplicativity)}\\
&\qquad\qquad + \|\boldsymbol{R}^{(L)}[r,:]\|_2\|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_2 + \|\boldsymbol{H}^{(L)}[t,:]\|_2\|\boldsymbol{V}_r^{\langle j\rangle}\|_2\Big)\\
&\leq (2\Phi_L + \Lambda_L)\, s_{L+1} \qquad \text{(definitions of } \Phi_l, \Lambda_L, \text{ and } s_{L+1}\text{)}\\
&\leq (2\eta_L + \|\boldsymbol{X}_{\text{rel}}\|_2)\, s_{L+1}\prod_{i=1}^{L} s_i \leq \zeta_L s^{L+1} < \frac{\gamma}{2}\\
&\to s < \left(\frac{\gamma}{2\zeta_L}\right)^{\frac{1}{L+1}}
\end{aligned}$$

then Theorem 4.4 trivially holds since $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) = \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) = 1$ when $\|f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{2}$ holds for all $(h,r,t)\in\mathcal{E}$.

Also, if

$$\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6N_{\mathbf{w}}(L+1)^2 s^{2L+2}\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} + \frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \geq \sqrt{\left(\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\right)\left(\frac{4s^{2L+2}\zeta_L^2}{\gamma^2}\right)} > 1$$

$$\to s > \left(\frac{\gamma}{2\zeta_L}\sqrt{\frac{|\widehat{\mathcal{E}}|}{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}}\right)^{\frac{1}{L+1}}$$

then Theorem 4.4 holds regardless of the value of $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ and $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ because the value of the loss function cannot exceed 1.

Therefore, we only need to consider s in range

$$\left(\frac{\gamma}{2\zeta_L}\right)^{\frac{1}{L+1}} \leq s \leq \left(\frac{\gamma}{2\zeta_L}\sqrt{\frac{|\widehat{\mathcal{E}}|}{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}}\right)^{\frac{1}{L+1}} \tag{14}$$

We also need to check whether the assumption $\ddot{s} \leq \frac{1}{L+2}s$ holds in this range. Note that if $\frac{\gamma}{4e^2(L+1)\zeta_L\left(\frac{L+1}{L+2}s\right)^L} \leq \frac{1}{L+2}s$ holds, then the assumption also holds since $\ddot{s} = \frac{\gamma}{4e^2(L+1)\mathring{s}^L\zeta_L} \leq \frac{\gamma}{4e^2(L+1)\zeta_L\left(\frac{L+1}{L+2}s\right)^L}$. With a simple calculation, we get

$$s^{L+1} \geq \frac{(L+2)\gamma}{4e^2(L+1)\zeta_L\left(\frac{L+1}{L+2}\right)^L}$$

The above inequality holds if s is in the range of Eq. (14), since

$$\frac{(L+2)\gamma}{4e^2(L+1)\zeta_L\left(\frac{L+1}{L+2}\right)^L} = \frac{\gamma}{4e^2\zeta_L}\left(1+\frac{1}{L+1}\right)^{L+1} \leq \frac{\gamma}{4e\zeta_L} \leq \frac{\gamma}{2\zeta_L} \leq s^{L+1} \qquad \left(\left(1+\frac{1}{x}\right)^x \leq e, \forall x \geq 0\right)$$

Therefore, we only need to consider Eq. (14) because otherwise Theorem 4.4 holds regardless of the choice of σ. While Eq. (13) holds with probability $1-\delta$, it only holds for s such that $\frac{L+1}{L+2}s \leq \mathring{s} \leq \frac{L+3}{L+2}s$. To make Eq.(13) hold for all s in range Eq. (14), we need to select multiple $\mathring{s}$ so that any s in range Eq. (14) can be covered. By assuming that $|s-\mathring{s}| \leq \frac{1}{L+2}\left(\frac{\gamma}{2\zeta_L}\right)^{\frac{1}{L+1}} \leq \frac{1}{L+2}s$, we can calculate the number of $\mathring{s}$ we need to consider, i.e., the size of covering C, by dividing the length of the range of s in Eq. (14) by the length of each cover, i.e., $\frac{2}{L+2}\left(\frac{\gamma}{2\zeta_L}\right)^{\frac{1}{L+1}}$. Let $|C|$ denote the size of covering C. By simple division, we get $|C| = \frac{(L+2)}{2}\left(\left(\sqrt{\frac{1}{|\widehat{\mathcal{E}}|}-\frac{1}{|\mathcal{E}|}}\right)^{-\frac{1}{L+1}}-1\right)$. Using Bernoulli's inequality, we can conclude that the probability of Eq. (13) holding simultaneously for $|C|$ choices of $\mathring{s}$ is $1-|C|\delta$. Therefore,

$$\begin{aligned}\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6N_{\mathbf{w}}(L+1)^2s^{2L+2}\zeta_L^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2}+\frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)|C|}{\delta}\right]}\\ \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}}L^2\zeta_L^2 s^{2L}d\ln(N_{\mathbf{w}}d)}{\gamma^2}+\ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right)\end{aligned}$$

holds with probability of $1-|C|\cdot\frac{\delta}{|C|} = 1-\delta$ regardless of s. □

Theorem 4.5 (Generalization Bound for ReED with Semantic Matching Decoder). *For any $L \geq 0$, let $f_{\mathbf{w}} : \mathcal{V}\times\mathcal{R}\times\mathcal{V} \to \mathbb{R}^2$ be a triplet classifier designed by the combination of the RAMP encoder with L-layers in Definition 3.1 and the SM decoder in Definition 3.3. Let k_r be the maximum of the infinity norms for all possible $\boldsymbol{S}_r^{(l)}$ in the RAMP encoder. Then, for any $\delta, \gamma > 0$, with probability at least $1-\delta$ over a training triplet set $\widehat{\mathcal{E}}$ (such that $20 \leq |\widehat{\mathcal{E}}| \leq |\mathcal{E}|-20$) sampled without replacement from the full triplet set $\mathcal{E}$, for any $\mathbf{w}$, we have*

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}}L^2\eta_L^4 s^{4L}d\ln(N_{\mathbf{w}}d)}{\gamma^2}+\ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right)$$

where $\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|) = 3\sqrt{|\widehat{\mathcal{E}}|(1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|})}\ln|\widehat{\mathcal{E}}|$, $N_{\mathbf{w}} = 2|\mathcal{R}|L + 2|\mathcal{R}| + 2L$, $\eta_L = \tau^L\|\boldsymbol{X}_{\mathrm{ent}}\|_2 + \kappa\|\boldsymbol{X}_{\mathrm{rel}}\|_2\sum_{i=0}^{L-1}\tau^i$, $d = \max(\max_{0\leq l\leq L}(d_l), \max_{0\leq l\leq L}(d'_l))$, $\tau = C_\phi + \kappa$, $\kappa = C_\phi C_\rho C_\psi\sum_{r\in\mathcal{R}}k_r$, $s_{L+1} = \max_{r,j}\|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_F$, $s_l = \max(\|\boldsymbol{W}_0^{(l)}\|_F, \|\boldsymbol{U}_0^{(l)}\|_F, \max_r\|\boldsymbol{W}_r^{(l)}\|_F, \max_r\|\boldsymbol{U}_r^{(l)}\|_F)$ for $l \in \{1,2,\ldots,L\}$, and $s = \max_{1\leq l\leq L+1}(s_l)$.

Proof. We derive Theorem 4.5 from Theorem 4.3 where we construct a posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ by adding random perturbations $\ddot{\mathbf{w}}$ to $\mathbf{w}$. Following [24; 35], we set the prior distribution $\mathcal{P}$ as $\mathcal{N}\left(\mathbf{0}_{n_\mathbf{w}}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w}\times n_\mathbf{w}}\right)$ and the posterior distribution $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ as $\mathcal{N}\left(\mathbf{w}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w}\times n_\mathbf{w}}\right)$ where $n_\mathbf{w}$ is the size of $\mathbf{w}$. We first compute $\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_\mathbf{w}(h,r,t)\|_\infty$, which we call the perturbation bound, so that we can calculate the standard deviation σ of the prior distribution that satisfies $\mathbb{P}\left(\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_\mathbf{w}(h,r,t)\|_\infty < \frac{\gamma}{4}\right) > \frac{1}{2}$. Afterwards, we calculate the KL divergence of $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ from $\mathcal{P}$ using σ and substitute the KL divergence term in Theorem 4.3 with our data and model-related terms, which finishes the proof.

Perturbation bound of ReED with semantic matching decoder First, we compute the perturbation bound, $\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_\mathbf{w}(h,r,t)\|_\infty$, and find σ that makes $\mathbb{P}(\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t)$-$f_\mathbf{w}(h,r,t)\|_\infty < \frac{\gamma}{4}) > \frac{1}{2}$ true. Let $\ddot{\boldsymbol{W}}$ denote a perturbation (also called noise) matrix added to the original weight matrix $\boldsymbol{W}$. As a result, we have $\widetilde{\boldsymbol{W}} = \boldsymbol{W} + \ddot{\boldsymbol{W}}$ where $\widetilde{\boldsymbol{W}}$ is a perturbed weight matrix. Also, let $\widetilde{\boldsymbol{H}}^{(l)}$ and $\widetilde{\boldsymbol{R}}^{(l)}$ denote the outputs of the perturbed model at the l-th layer. Each element of $\ddot{\boldsymbol{W}}$ is an i.i.d. element drawn from $\mathcal{N}\left(0, \sigma^2\right)$. Assume that the maximum of the Frobenius norms of the noise matrices is $\ddot{s}$. That is,

$$\ddot{s} = \max_l \left(\left\|\ddot{\boldsymbol{W}}_0^{(l)}\right\|_F, \left\|\ddot{\boldsymbol{U}}_0^{(l)}\right\|_F, \max_r \left\|\ddot{\boldsymbol{W}}_r^{(l)}\right\|_F, \max_r \left\|\ddot{\boldsymbol{U}}_r^{(l)}\right\|_F, \max_{r,j} \left\|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\right\|_F \right)$$

Now, let us calculate the perturbation bound.

$$\begin{aligned}
&\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_\mathbf{w}(h,r,t)\|_\infty \\
&= \max_{(h,r,t)\in\mathcal{E}} \max_{j\in\{0,1\}} |\widetilde{\boldsymbol{H}}^{(L)}[h,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} \left(\widetilde{\boldsymbol{H}}^{(L)}[t,:]\right)^\top - \boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} \left(\boldsymbol{H}^{(L)}[t,:]\right)^\top |
\end{aligned}$$

Recall that $\Phi_l = \max_v \|\boldsymbol{H}^{(l)}[v,:]\|_2$ and $\Psi_l = \max_v \|\widetilde{\boldsymbol{H}}^{(l)}[v,:] - \boldsymbol{H}^{(l)}[v,:]\|_2$.

For any $(h,r,t)\in\mathcal{E}$,

$$\begin{aligned}
&\left|\widetilde{\boldsymbol{H}}^{(L)}[h,:]\widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle} \left(\widetilde{\boldsymbol{H}}^{(L)}[t,:]\right)^\top - \boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle} \left(\boldsymbol{H}^{(L)}[t,:]\right)^\top\right| \\
=&\Big|\left(\widetilde{\boldsymbol{H}}^{(L)}[h,:] - \boldsymbol{H}^{(L)}[h,:] + \boldsymbol{H}^{(L)}[h,:]\right)\left(\overline{\boldsymbol{U}}_r^{\langle j\rangle} + \ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\right)\left(\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:] + \boldsymbol{H}^{(L)}[t,:]\right)^\top \\
&-\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle}\left(\boldsymbol{H}^{(L)}[t,:]\right)^\top\Big| \quad \text{(definition of } \widetilde{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\text{)} \\
=&\Big|\left(\widetilde{\boldsymbol{H}}^{(L)}[h,:] - \boldsymbol{H}^{(L)}[h,:]\right)\overline{\boldsymbol{U}}_r^{\langle j\rangle}\left(\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:] + \boldsymbol{H}^{(L)}[t,:]\right)^\top \\
&+ \boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle j\rangle}\left(\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:]\right)^\top \\
&+ \left(\widetilde{\boldsymbol{H}}^{(L)}[h,:] - \boldsymbol{H}^{(L)}[h,:] + \boldsymbol{H}^{(L)}[h,:]\right)\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\left(\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:] + \boldsymbol{H}^{(L)}[t,:]\right)^\top\Big| \\
\leq&\|\widetilde{\boldsymbol{H}}^{(L)}[h,:] - \boldsymbol{H}^{(L)}[h,:]\|_2\|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_2\left(\|\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:]\|_2 + \|\boldsymbol{H}^{(L)}[t,:]\|_2\right) \\
&+ \|\boldsymbol{H}^{(L)}[h,:]\|_2\|\overline{\boldsymbol{U}}_r^{\langle j\rangle}\|_2\|\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:]\|_2 \quad \text{(sub-additivity, sub-multiplicativity)} \\
&+ \left(\|\widetilde{\boldsymbol{H}}^{(L)}[h,:] - \boldsymbol{H}^{(L)}[h,:]\|_2 + \|\boldsymbol{H}^{(L)}[h,:]\|_2\right)\left\|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\right\|_2\left(\|\widetilde{\boldsymbol{H}}^{(L)}[t,:] - \boldsymbol{H}^{(L)}[t,:]\|_2 + \|\boldsymbol{H}^{(L)}[t,:]\|_2\right) \\
\leq&\Psi_L\left(\Psi_L + 2\Phi_L\right)s_{L+1} + \left(\Psi_L + \Phi_L\right)^2\ddot{s} \qquad \text{(definitions of } \Psi_L, \Phi_L, \text{ and } s_{L+1}\text{)}
\end{aligned}$$

Therefore,

$$\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_\mathbf{w}(h,r,t)\|_\infty \leq \Psi_L(\Psi_L + 2\Phi_L)s_{L+1} + (\Psi_L + \Phi_L)^2\ddot{s}$$

Note that we can calculate the upper bounds of Φ_L and Ψ_L using Eq. (9) and Eq. (11), respectively. Then,

$$
\begin{aligned}
&\max_{(h,r,t)\in\mathcal{E}} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty \\
&\leq \Psi_L(\Psi_L + 2\Phi_L)s_{L+1} + (\Psi_L + \Phi_L)^2 \ddot{s} \\
&\leq \eta_L \left(\prod_{i=1}^{L}(s_i + \ddot{s}) - \prod_{i=1}^{L} s_i\right)\left(\eta_L\left(\prod_{i=1}^{L}(s_i + \ddot{s}) - \prod_{i=1}^{L} s_i\right) + 2\eta_L \prod_{i=1}^{L} s_i\right) s_{L+1} \\
&\quad + \left(\eta_L\left(\prod_{i=1}^{L}(s_i + \ddot{s}) - \prod_{i=1}^{L} s_i\right) + \eta_L \prod_{i=1}^{L} s_i\right)^2 \ddot{s} \\
&= \eta_L^2 \left(\left(\prod_{i=1}^{L}(s_i + \ddot{s})\right)^2 - \left(\prod_{i=1}^{L} s_i\right)^2\right) s_{L+1} + \eta_L^2 \left(\prod_{i=1}^{L}(s_i + \ddot{s})\right)^2 \ddot{s} \\
&= \eta_L^2 \left(\left(\left(\prod_{i=1}^{L}(s_i + \ddot{s})\right)^2 - \left(\prod_{i=1}^{L} s_i\right)^2\right) s_{L+1} + \left(\prod_{i=1}^{L}(s_i + \ddot{s})\right)^2 \ddot{s}\right) \\
&\leq \eta_L^2 \left(\left((s+\ddot{s})^{2L} - s^{2L}\right) s + (s+\ddot{s})^{2L}\ddot{s}\right) \qquad \text{(definition of } s\text{)} \\
&= \eta_L^2 \left((s+\ddot{s})^{2L+1} - s^{2L+1}\right)
\end{aligned} \tag{15}
$$

Generalization bound of ReED with semantic matching decoder Recall that we set the prior distribution $\mathcal{P}$ as $\mathcal{N}\left(\mathbf{0}_{n_\mathbf{w}}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w}\times n_\mathbf{w}}\right)$. Since the distribution of the perturbed parameters $\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}$ is $\mathcal{N}\left(\mathbf{w}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w}\times n_\mathbf{w}}\right)$, the distribution of the perturbation is $\mathcal{N}\left(\mathbf{0}_{n_\mathbf{w}}, \sigma^2 \boldsymbol{I}_{n_\mathbf{w}\times n_\mathbf{w}}\right)$. For the perturbation matrices that follow the normal distribution, we can derive the following by replacing the standard normal variable in Corollary 4.2. in Tropp [48] with a random variable following the normal distribution having the mean of 0 and the variance of σ^2.

$$
\begin{aligned}
&\mathbb{P}\left(\|\ddot{\boldsymbol{W}}_0^{(l)}\|_2 \geq \ddot{s}\right) \leq (d_{l-1} + d_l)e^{-\ddot{s}^2/(2\max(d_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)} \\
&\mathbb{P}\left(\|\ddot{\boldsymbol{U}}_0^{(l)}\|_2 \geq \ddot{s}\right) \leq (d'_{l-1} + d'_l)e^{-\ddot{s}^2/(2\max(d'_{l-1},d'_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)} \\
&\mathbb{P}\left(\|\ddot{\boldsymbol{W}}_r^{(l)}\|_2 \geq \ddot{s}\right) \leq (d_{l-1} + d_l)e^{-\ddot{s}^2/(2\max(d_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)} \\
&\mathbb{P}\left(\|\ddot{\boldsymbol{U}}_r^{(l)}\|_2 \geq \ddot{s}\right) \leq (d'_{l-1} + d_l)e^{-\ddot{s}^2/(2\max(d'_{l-1},d_l)\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)} \\
&\mathbb{P}\left(\|\ddot{\overline{\boldsymbol{U}}}_r^{\langle j\rangle}\|_2 \geq \ddot{s}\right) \leq 2d_L e^{-\ddot{s}^2/(2d_L\sigma^2)} \leq 2de^{-\ddot{s}^2/(2d\sigma^2)}
\end{aligned}
$$

Using Bernoulli's inequality, we can derive that the probability of all perturbation matrices having the spectral norm less than $\ddot{s}$ is greater than or equal to $1 - 2N_\mathbf{w} de^{-\ddot{s}^2/(2d\sigma^2)}$, where $N_\mathbf{w} = 2\cdot L + 2\cdot|\mathcal{R}|L + 2|\mathcal{R}| = 2\cdot|\mathcal{R}|L + 2|\mathcal{R}| + 2L$ is the number of perturbation matrices.

To satisfy the condition of Theorem 4.3, we set $2N_\mathbf{w} de^{-\ddot{s}^2/(2d\sigma^2)} = 1/2$. Then, we get $\ddot{s} = \sigma\sqrt{2d\ln(4N_\mathbf{w} d)}$. Since the prior is independent of the learned parameters $\mathbf{w}$, we cannot directly use s to formulate σ. Therefore, we approximate s with $\mathring{s}$ in the following range.

$$
|s - \mathring{s}| \leq \frac{1}{2L+2}s \implies \frac{2L+1}{2L+2}s \leq \mathring{s} \leq \frac{2L+3}{2L+2}s
$$

Additionally, we assume $\ddot{s} \le \frac{1}{2L+2}s$. Then, if

$$\begin{aligned}
\max_{(h,r,t)} \|f_{\mathbf{w}+\ddot{\mathbf{w}}}(h,r,t) - f_{\mathbf{w}}(h,r,t)\|_\infty &\le \eta_L^2 \left((s+\ddot{s})^{2L+1} - s^{2L+1}\right) && \text{(Eq.(15))} \\
&\le \eta_L^2 \ddot{s}(2L+1)(s+\ddot{s})^{2L} && (0 \le s \le s+\ddot{s}) \\
&\le \eta_L^2 \ddot{s}(2L+1)s^{2L}\left(1+\frac{1}{2L+2}\right)^{2L} && \left(\ddot{s} \le \frac{1}{2L+2}s\right) \\
&\le \eta_L^2 \ddot{s}(2L+1)\left(\frac{2L+2}{2L+1}\mathring{s}\right)^{2L}\left(1+\frac{1}{2L+2}\right)^{2L} && \text{(range of } \mathring{s}) \\
&= \eta_L^2 \ddot{s}(2L+1)\mathring{s}^{2L}\left(1+\frac{2}{2L+1}\right)^{2L} \\
&\le \eta_L^2 \ddot{s}(2L+1)\mathring{s}^{2L}e^2 \le \frac{\gamma}{4} && \left(\left(1+\frac{1}{x}\right)^x \le e, \forall x \ge 0\right)
\end{aligned}$$

is satisfied, we meet the condition of Theorem 4.3. With $\ddot{s} = \sigma\sqrt{2d\ln(4N_{\mathbf{w}}d)}$, we have

$$\begin{aligned}
\ddot{s} = \sigma\sqrt{2d\ln(4N_{\mathbf{w}}d)} &\le \frac{\gamma}{4e^2(2L+1)\eta_L^2\mathring{s}^{2L}} \\
\rightarrow \sigma &\le \frac{1}{\sqrt{2d\ln(4N_{\mathbf{w}}d)}}\left(\frac{\gamma}{4e^2(2L+1)\eta_L^2\mathring{s}^{2L}}\right)
\end{aligned}$$

By setting $\sigma = \frac{1}{\sqrt{2d\ln(4N_{\mathbf{w}}d)}}\left(\frac{\gamma}{4e^2(2L+1)\eta_L^2\mathring{s}^{2L}}\right)$, we can calculate $D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P})$.

$$\begin{aligned}
D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) &= \frac{\|\mathbf{w}\|_2^2}{2\sigma^2} && \text{(KL divergence between two normal distributions)} \\
&= \frac{2d\ln(4N_{\mathbf{w}}d)}{2\left(\frac{\gamma}{4e^2(2L+1)\eta_L^2\mathring{s}^{2L}}\right)^2}\left(\sum\|\boldsymbol{W}\|_F^2\right) \\
&\le \frac{(4e^2(2L+1)\eta_L^2\mathring{s}^{2L})^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2}N_{\mathbf{w}}s^2 \\
&\le \frac{N_{\mathbf{w}}s^2\left(4e^2(2L+1)\eta_L^2\left(\frac{2L+3}{2L+2}s\right)^{2L}\right)^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} && \text{(range of } \mathring{s}) \\
&\le \frac{N_{\mathbf{w}}s^2(4e^3(2L+1)\eta_L^2 s^{2L})^2 d\ln(4N_{\mathbf{w}}d)}{\gamma^2} && \left(\left(1+\frac{1}{x}\right)^x \le e, \forall x \ge 0\right) \\
&= \frac{16e^6 N_{\mathbf{w}}(2L+1)^2\eta_L^4 s^{4L+2} d\ln(4N_{\mathbf{w}}d)}{\gamma^2}
\end{aligned}$$

From Theorem 4.3, we get

$$\begin{aligned}
\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) &\le \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{2|\widehat{\mathcal{E}}|}\left[2D_{KL}(\mathcal{Q}_{\mathbf{w}+\ddot{\mathbf{w}}}\|\mathcal{P}) + \ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \\
&\le \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6 N_{\mathbf{w}}(2L+1)^2\eta_L^4 s^{4L+2} d\ln(4N_{\mathbf{w}}d)}{\gamma^2} + \frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \qquad (16)
\end{aligned}$$

Now, let us find some range of s such that Theorem 4.5 trivially holds. First, if

$$\|f_{\mathbf{w}}(h,r,t)\|_\infty \leq \max\left(\left|\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle 1\rangle}\left(\boldsymbol{H}^{(L)}[t,:]\right)^\top\right|, \left|\boldsymbol{H}^{(L)}[h,:]\overline{\boldsymbol{U}}_r^{\langle 2\rangle}\left(\boldsymbol{H}^{(L)}[t,:]\right)^\top\right|\right)$$
$$\leq \Phi_L^2 s_{L+1} \leq s\eta_L^2\left(\prod_{i=1}^{L} s_i\right)^2 \leq \eta_L^2 s^{2L+1} < \frac{\gamma}{2}$$
$$\to s < \left(\frac{\gamma}{2\eta_L^2}\right)^{\frac{1}{2L+1}}$$

then Theorem 4.5 trivially holds since $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) = \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) = 1$ when $\|f_{\mathbf{w}}(h,r,t)\|_\infty < \frac{\gamma}{2}$ holds for all $(h,r,t) \in \mathcal{E}$. Also, if

$$\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6 N_{\mathbf{w}}(2L+1)^2\eta_L^4 s^{4L+2} d\ln(4N_{\mathbf{w}}d)}{\gamma^2} + \frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]} \geq \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\frac{4s^{4L+2}\eta_L^4}{\gamma^2}} > 1$$
$$\to s > \left(\frac{\gamma}{2\eta_L^2}\sqrt{\frac{|\widehat{\mathcal{E}}|}{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}}\right)^{\frac{1}{2L+1}}$$

then Theorem 4.5 holds regardless of the value of $\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}})$ and $\mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}})$ because the value of the loss cannot exceed 1. Therefore, we only need to consider s in range

$$\left(\frac{\gamma}{2\eta_L^2}\right)^{\frac{1}{2L+1}} \leq s \leq \left(\frac{\gamma}{2\eta_L^2}\sqrt{\frac{|\widehat{\mathcal{E}}|}{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}}\right)^{\frac{1}{2L+1}} \tag{17}$$

We also need to check whether the assumption $\ddot{s} \leq \frac{1}{2L+2}s$ holds in this range. Note that if $\frac{\gamma}{4e^2(2L+1)\eta_L^2\left(\frac{2L+1}{2L+2}s\right)^{2L}} \leq \frac{1}{2L+2}s$ holds, then the assumption also holds since $\ddot{s} = \frac{\gamma}{4e^2(2L+1)\eta_L^2\mathring{s}^{2L}} \leq \frac{\gamma}{4e^2(2L+1)\eta_L^2\left(\frac{2L+1}{2L+2}s\right)^{2L}}$. With a simple calculation, we get

$$s^{2L+1} \geq \frac{(2L+2)\gamma}{4e^2(2L+1)\eta_L^2\left(\frac{2L+1}{2L+2}\right)^{2L}}$$

The above inequality holds if s is in the range of Eq. (17) since

$$\frac{(2L+2)\gamma}{4e^2(2L+1)\eta_L^2\left(\frac{2L+1}{2L+2}\right)^{2L}} = \frac{\gamma}{4e^2\eta_L^2}\left(1+\frac{1}{2L+1}\right)^{2L+1} \leq \frac{\gamma}{4e\eta_L^2} \leq \frac{\gamma}{2\eta_L^2} \leq s^{2L+1} \qquad \left(\left(1+\frac{1}{x}\right)^x \leq e, \forall x \geq 0\right)$$

Therefore, we only need to consider Eq. (17) because otherwise Theorem 4.5 holds regardless of the choice of σ. While Eq. (16) holds with probability $1-\delta$, it only holds for s such that $\frac{2L+1}{2L+2}s \leq \mathring{s} \leq \frac{2L+3}{2L+2}s$. To make Eq. (16) hold for all s in range Eq. (17), we need to select multiple $\mathring{s}$ so that any s in range Eq. (17) can be covered. By assuming that $|s-\mathring{s}| \leq \frac{1}{2L+2}\left(\frac{\gamma}{2\eta_L^2}\right)^{\frac{1}{2L+1}} \leq \frac{1}{2L+2}s$, we can calculate the number of $\mathring{s}$ we need to consider, i.e., the size of covering C, by dividing the length of the range of s in Eq. (17) by the length of each cover, i.e., $\frac{2}{2L+2}\left(\frac{\gamma}{2\eta_L^2}\right)^{\frac{1}{2L+1}}$. Let $|C|$ denote the size of covering C. By simple division, we get $|C| = \frac{(2L+2)}{2}\left(\left(\sqrt{\frac{1}{|\widehat{\mathcal{E}}|}-\frac{1}{|\mathcal{E}|}}\right)^{-\frac{1}{2L+1}} - 1\right)$. Using Bernoulli's inequality, we can

Table 4: Dataset Statistic

	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\widehat{\mathcal{E}}\rvert$	$\lvert\mathcal{E}\rvert$
FB15K237	1,496	179	41,873	52,318
CoDEx-M	2,684	42	17,951	22,224
UMLS-43	133	43	10,174	12,732

conclude that the probability of Eq. (16) holding simultaneously for $|C|$ choices of $\mathring{s}$ is $1-|C|\delta$. Therefore,

$$\begin{aligned}\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{16e^6 N_{\mathbf{w}}(2L+1)^2\eta_L^4 s^{4L+2} d\ln(4N_{\mathbf{w}}d)}{\gamma^2} + \frac{1}{2}\ln\frac{4\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)|C|}{\delta}\right]} \\ \leq& \mathcal{L}_{\gamma,\widehat{\mathcal{E}}}(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{1-\frac{|\widehat{\mathcal{E}}|}{|\mathcal{E}|}}{|\widehat{\mathcal{E}}|}\left[\frac{N_{\mathbf{w}}L^2\eta_L^4 s^{4L} d\ln(N_{\mathbf{w}}d)}{\gamma^2} + \ln\frac{\theta(|\widehat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right)\end{aligned}$$

holds with probability of $1-|C|\cdot\frac{\delta}{|C|}=1-\delta$. □

E. Experimental Details

We conduct experiments on three real-world knowledge graphs: FB15K237 [47], CoDEx-M [40], and UMLS-43 [7; 27], shown in Table 4. We generate smaller versions of FB15K237 and CoDEx-M via graph sampling [46] for ease of analysis. We create smaller versions of FB15K237 [47] and CoDEx-M [40] using a standard graph sampling [46] and consider them as fully observed knowledge graphs G for easier analysis. Specifically, we randomly sample five seed entities for FB15K237 and ten seed entities for CoDEx-M. From the seed entities, we randomly sample 30 neighboring entities per hop for two hops. Then, we take all sampled entities and the triplets between them. While CoDEx-M contains negative triplets (i.e., false triplets needed for training a triplet classifier), FB15K237 and UMLS-43 do not include negative triplets. For these datasets, we create negative triplets by corrupting either a head or a tail entity of each positive triplet, following Socher et al. [44].

We set $d_1 = d_2 = \cdots = d_L = d_{L+1}$ for all datasets. We use $d_1 = 96$ for FB15K237, $d_1 = 64$ for CoDEx-M, and $d_1 = 48$ for UMLS-43. For RAMP+TD, we set the learning rate to be 0.0003 on FB15K237, 0.0005 for CoDEx-M, and 0.0002 for UMLS-43. For RAMP+SM, we set the learning rate to be 0.0005 for all datasets. We set the margin of the margin loss for FB15K237 and CoDEx-M to be 0.5, and 0.75 for UMLS-43 and run all models for 2,000 epochs.

In our implementation of ReED, we use the Adam optimizer [20]. When implementing ReED, we used python 3.8 and PyTorch 1.12.1 with cudatoolkit 11.3. We run all our experiments using NVIDIA GeForce RTX 2080 Ti. We repeat each experiment ten times with the random seeds: 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90. Our code and data are available at `https://github.com/bdi-lab/ReED` where more details about the experiments are explained in the README file.